The Effect of Copy Number Variation on Human Phenotypes

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

BY

Majid Ibrahim Alsagabi

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Ahmed H. Tewfik, Adviser
September 2012

Abstract

The human DNA copy number variation (DCV) has been proven to be correlated to abnormal traits and features in human beings. The genomic hybridization experiment is a powerful biological tool to measure the level of the DNA copy number in thousands or millions of genomic sites simultaneously. The experiment is subject to large amounts of noise and a high level of uncertainty about the biological meaning of its measurements.

The existing methods to detect the DCV are based on the two-channel approach which consists of test and reference samples. Most of the methods are ill conditioned for large data sets because of their complexity and sophisticated approaches. Furthermore, they fall short of achieving an acceptable sensitivity or they generate large amounts of false calls. The first part of this thesis explores the existing methods and presents four new models to simplify the solution. The four models are based on Band-Pass Wavelet Transform, Uncovered Markov Model, the Uniformly Most Powerful Test, and the Maximum Likelihood Estimator. The four models achieve the highest sensitivity, lowest false alarm rate, and the least complexity of all models.

The second part of the thesis presents a novel model for DCV detection using a single-channel approach. The model is based on the concept of sensor networks which can be used to analyze the DNA samples from one or two channels. The model comprises three normalization techniques to remove the non-biological bias from the measurements. Then, it estimates the true distribution of the normal measurements by isolating their distribution from the heterogeneous mixture. The complexity of calculating the probability of the average error is overcome by using the saddle-point approximation and the log-lattice design. The accuracy of the saddle-point approximation is proven for both the two-channel and the single-channel approaches in homogenous and non-homogenous environments. The analysis includes both simulated and real-world datasets and it explores the recurrent DCV in large populations using the International Hapmap Project Datasets. The end of the second part of the thesis demonstrates the stationarity of the hybridization experiment and shows its impact on reducing the complexity of the analysis.

The third part of the thesis investigates patterns of the DNA copy number variations. The human genetic network is a quite complex system where hundreds, or even thousands, of DNA segments interact internally with each other directly or indirectly to control all the body's functions. A bottom-up subspace-clustering algorithm is presented to reveal the biological signature of two studied phenotypes: Autism, and the lethal castration-resistant prostate cancer.

Contents

Abstract i

List of Tables vi

List of Figures vii

1 Introduction 1

1.1 Overview ……. 1

1.2 Motivation ……. 4

1.3 Contributions and Chapters Description ……. 5

2 DNA Copy Number Profiling Using Two-Channel Approaches 7

2.1 Genomic Hybridization Experiment ……. 7

2.2 Data Modeling ……. 9

2.3 Related Work ……. 10

2.3.1 Finite Impulse Response Filters ……. 11

2.3.2 Hidden Markov Models ……. 12

2.3.3 Maximum Likelihood Estimators ……. 13

2.3.4 Neyman-Pearson Theory ……. 14

2.4 Our contribution ……. 15

2.4.1 Band-Pass Wavelet Transform ……. 15

2.4.2 Uncovered Markov Model ……. 16

2.4.3 Truncated Likelihood Ratio Test ……. 18

2.4.4 Minimum Interval Score ……. 20

2.5 Comprehensive Comparison Using Real-World Data ……. 23

2.5.1 Finely Tiled Arrays ……. 23

2.5.2 QPCR Test ……. 25

2.5.3 Sensitivity Versus False Alarm ……. 25

2.5.4 Data Modeling ……. 26

2.5.5 Results and Discussion ……. 27

2.6 Complexity of BPWT, UMM, TLRT, and MIS ……..……. 32
2.7 Reproducibility of ROC curves ……..…… 33
2.8 Conclusions ……..…… 38

3 A Statistical Model For Genomic Hybridization Experiments ……..…... 39
3.1 Introduction ……..……. 39
3.2 Genome-Wide Human SNP Array 6.0 – Affymetrix ……..…… 41
3.3 Quantile-Based Perfectly-Isolated Model QPI ……..…….. 45
3.3.1 Related Work ……..……. 45
3.3.2 QPI Model ……..……..… 51
3.3.3 Modeling the intensity distribution of gains and losses 55
3.4 Removal of Systematic Bias ……..……...… 57
3.4.1 Introduction to Scanner bias ……..……...…. 58
3.4.2 Related work ……..……. 59
3.4.3 Universal-Threshold Adjustment (UTA) Algorithm …. 60
3.5 Removal of GC-Content Bias ……..……..... 62
3.5.1 Introduction and related work ……..……. 62
3.5.2 GCNORM model ……..……..…. 63
3.6 Removal of Fragment Length Bias ……..…….. 65
3.6.1 Introduction and related work ……..……. 65
3.6.2 FLNORM model ……..……...…. 67
68
3.7 Results and Discussion ……..……..….
3.7.1 The Data of Hapmap project ……..……... 68
3.7.2 Results of the UTA algorithm ……..……. 69
3.7.3 Results of GCNORM ……..……. 73
3.7.4 Results of FLNORM ……..……. 75
3.8 Microarrays Stationarity ……..……..… 75
3.9 Conclusions ……..……...…… 78

4 Sensor Network Approach for DNA Copy Number Microarrays ……. 79
4.1 Sensor Networks for DNA Microarrays ……..…….. 81
4.2 Saddle-Point Approximation for Heterogeneous Environments ... 83
4.3 Log-Lattice Quantizer ……..…….…….…… 85
4.4 The accuracy of the saddle-point approximation in Microarrays .. 87
4.5 The accuracy of the log-lattice lemma ……..…….…… 88
4.6 Experimental results of the sensor networks approach …….. 91
4.6.1 Two-channel approach ……..…….……... 91
4.6.2 The effect of the molecules number and network size .. 93
4.6.3 The stability and variability of the human genome …... 95
4.7 Conclusions……..…….…….…….…….…….…….…….. 99

5 Correlation Between Copy Number Variation and Human Diseases 100
5.1 Introduction ……..…….…….…….…….……... 101
2.2 Related Work ……..…….…….…….…….…… 102
5.3 Segregated-Based Clustering Algorithm SBC ……..…….…… 103
5.3.1 Collecting the set of candidate features ……..…….…… 104
5.3.2 Bottom-up approach ……..…….…….…… 105
5.4 Experimental Results On Autism ……..…….…….…… 105
5.4.1 Results and Conclusion ……..…….…….. 108
5.5 The Association of DNA Copy Number Variations with Prostate Cancer Therapy ……..…….…….…….…… 110
5.6 Conclusion ……..…….…….…….……... 110
6 Conclusion and Future Work 111
6.1 Two Channel Approaches ……..…….…….…….…… 111
6.2 Single Channel Approaches ……..…….…….…….…… 112
6.3 Subspace-Clustering Algorithms ……..…….…….…… 113

References 114

List of tables

2.1 QPCR sites for NA10851 versus NA15510 ……. 24

2.2 Additions and multiplications required by several algorithms …..…….. 32

2.3 Range of sensitivity and false alarm rate at selected tuning parameters . 37

3.1 Statistics of CN and SNP probes on the GWS6 ……...…… 43

5.1 Complexity of different techniques ……..….. 106

5.2 The size of the candidate features' set of different techniques ……..… 107

5.3 Comparative clustering results ……..…… 107

List of figures

1.1 DNA base composition 2
1.2 Types of variation of the DNA copy number 3
2.1 Filter banks of discrete and stationary wavelet transforms 12
2.2 SNR versus window size for multiple levels of false alarm 20
2.3 Numerical result of CBS 22
2.4 The autocorrelation of self-self experiment # 1..... 26
2.5 Q-Q plots of a self-self array versus a normal distribution 26
2.6 ROC curves of several poorly performing algorithms 27
2.7 ROC curves of 4 FIR filters 28
2.8 ROC curves of 4 Markov Models 28
2.9 ROC curves of 6 MLE models 29
2.10 ROC curves of 3 Neyman-Pearson models 30
2.11 ROC curves of our BPWT, TLRT, UMM, and MIS 31
2.12 AUC and residual for various algorithms 31
2.13 ROC curves for MIS, UMM, BPWT, and TLRT 33
2.14 Sensitivity of the tested arrays versus the training arrays 34
2.15 Variability of sensitivity under different tuning parameters of MIS ... 35
2.16 False alarm rates of the tested versus the training arrays 36
2.17 The variability of the FPR versus the tuning parameter in MIS 37
3.1 The intensity distribution of SNP probes 42
3.2 The probes' layout on a typical GWS6 chips 44
3.3 CN intensities histogram of sample NA18488 46
3.4 Cross correlation among "EPODE" sample of the IHP 47
3.5 Histogram of log2 ratios and normal distribution 50
3.6 Relative error of the closed forms of the quantiles 54

3.7 Typical images of Affymetrix 6.0 arrays …… 58
3.8 The effect of the GC-content on the intensity level …… 64
3.9 Intensity mean versus fragment length in StyI and NspI channels … 67
3.10 Intensity mean as a function of StyI and NspI channels ……...… 68
3.11 Three randomly selected samples from the Hapmap project …… 70
3.12 Edge detector using moving mean and median windows …… 70
3.13 Distributions of dark and bright areas before and after the normalization 71
3.14 Comparison of image's distributions before and after the normalization 72
3.15 Intensity bias versus the GC-content before and after the normalization 73
3.16 Result of GC-content normalizer in Partek Genomic Suite software … 74
3.17 Normalized intensity mean using FLNORM …… 74
3.18 Normalized intensity mean using PGS …… 74
3.19 Histograms to extract UT, SF, Δ, β, relative mean, and variance ……..... 76
4.1 Parallel fusion network …… 81
4.2 The average probability of error versus the number of nodes …… 89
4.3 Relative error of the P_e versus the number of nodes ……..… 89
4.4 The asymptotic relative efficiency versus the number of nodes ……. 90
4.5 ROC curves of TLRT, MIS, and the saddle-point approximation ……. 92
4.6 Optimal quantizer to detect the CNV in the two-channel microarrays …. 92
4.7 Optimized model to analyze two-channel arrays …….. 93
4.8 ROC curves for several values of shape parameter and network size ….. 94
4.9 Duplication and deletion rates in normal human genome ……. 96
4.10 Quantified duplication and deletion rates in normal human genome …. 97
4.11 Chromosomal duplication and deletion rates at 25%, 50% and 75% …. 98
4.12 Chromosomal duplications and deletion rates at 95% …….. 98
4.13 Chromosomes stability using 1% criteria …….. 99
5.1 Flowchart of SBC …….……. 106
5.2 Centroids of 5 Clusters in 13-D sup-space …….. 109

Chapter 1

Introduction

1.1 Overview

The deoxyribonucleic acid (DNA) is the main blueprint of all life forms except RNA viruses. Certain segments of the DNA, called genes, contain all the genetic information about the organism. Other parts of the DNA have minor structural or regulating functions, but the functions of most of the DNA parts are still unknown or not fully understood. The genes control the mechanism of all biological functions in the organism through the production of functional Ribonucleic acid (RNA) and protein.

The DNA consists of two long strands made up of phosphate groups and sugar connected to nucleobases. Each base from one strand is connected to a base on the other strand through hydrogen bonds. The two bases with the hydrogen bonds compose a base pair, while the base pair with the phosphate group and the sugar compose a nucleotide. There are only 4 types of bases in the genome: adenine (A), thymine (T), cytosine (C), and guanine (G). The base A pairs exclusively with T through two hydrogen bonds, and C pairs exclusively with G through three hydrogen bonds. The genetic code is stored in the sequence of these four bases along the strands. See figure 1.1. This sequence is the main factor that controls phenotypes, traits, and cellular activities. The differences within and among species are mainly due to the differences in the DNA sequences.

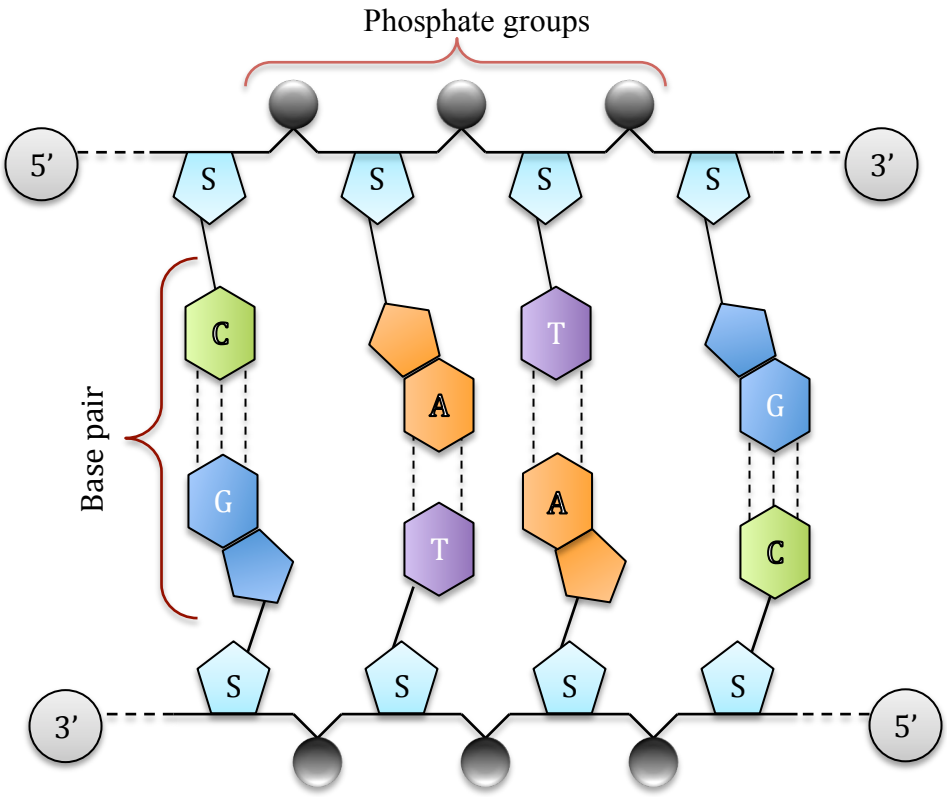


Figure 1.1: DNA base composition. S: sugar, A: adenine, T: thymine, C: cytosine, and G: guanine.

The human genome consists of about 3 billion base pairs divided into 46 chromosomes. 22 chromosomes are identical to 22 other chromosomes and the 22 pairs comprise the autosomes. The last pair includes the sex chromosomes: XX in females and XY in males. The human genome contains 20,000-25,000 genes covering only 1.5% of its total length. Although more than 99% of the human genome's sequence is identical in all people, no two individuals are identical. All differences among humans are caused by the differences in 1% of their genomes.

The variability of the genome occurs in two main forms: single nucleotide polymorphism (SNP) and copy number variation. The single nucleotide polymorphism occurs when a certain nucleotide differs between large groups of humans. For example,

a nucleotide A-T might convert to T-A, G-C, or C-G in at least 5% of all humans. The difference is limited to a single base called allele and fortunately, there are only two alleles for the vast majority of the common SNPs. There are approximately 3 million SNPs on the human genome which roughly represents 0.1% of its total length. The detection and the analysis of SNPs are beyond the scope of this thesis and will not be covered in our work.

The copy number variation occurs more frequently than SNPs and covers much longer portions of the genome. In the ideal case, the genome carries two copies of its exact sequence, one in each side of the paired autosomes, and only one copy of the sex chromosomes in males. The copy number variation is defined as any abnormality or amputation that occurs at any *section* of the genome. While the SNP affects only a single base, the copy number variation usually occurs at sections of length more than 1000 base pairs (bps). The main types of variation are duplication, deletion, insertion, and inversion. Figure 1.2 illustrates an example of each type. The illustrated variations are too short to be called copy number variation but we used them as examples because the space is limited. The total number of copies of a specific sequence changes to higher than normal in the duplication case, lower than normal in the deletion case, and it remains the same in the inversion and insertion cases.

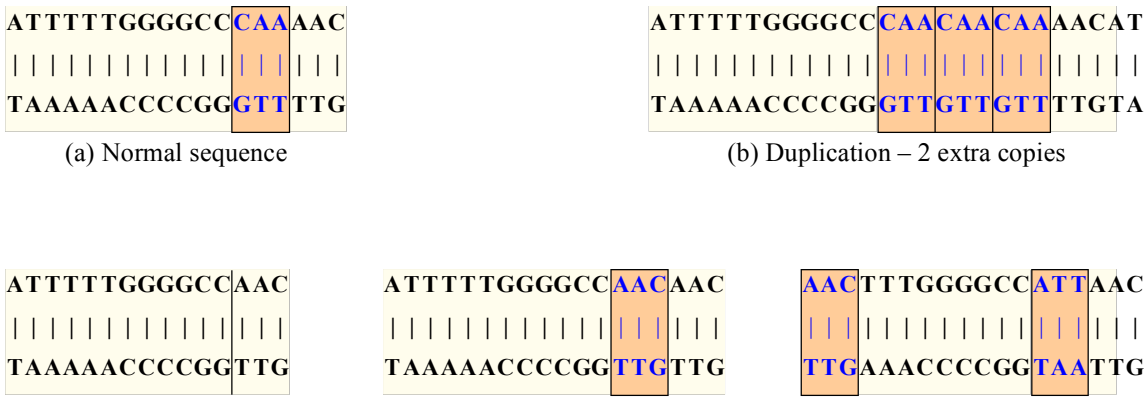


Figure 1.2: Types of variation of the DNA copy number

There are several methods to measure the amount of DNA copy number. One of the most efficient methods is the genomic hybridization process. The principal concept of the process is to dismantle the DNA sequence into short fragments and then to remove one the double strands from each fragment. The complementary sequence of each fragment is fabricated on a small chip called the array. The single-stranded fragments are called targets, while the fabricated sequences on the chip are called probes. There is a specific probe designed for each speicific target, and the binding process occurs based on the affinity between them. The resolution of the array is measured as the average number of probes within a fixed length of the genome and it varies drastically from one platform to another.

1.2 Motivation

The study of detecting the alteration in the DNA copy number (DCN) has drawn a lot of attention in the last decade. At the beginning, the work focused on observing the development and the progression of numerous types of cancers based on the genetic alterations. Recently, the field's scope extended to cover a wide spectrum of diseases that have been proven to be related to the DNA copy number variation (DCV). Several studies have reported independently that frequent gains and losses are widespread all over the human genome and they naturally cover up to 12% of the whole genome of typical normal humans [1]. Unlike the variation in cancerous tissues where the amputation starts at any stage of life, CNV in normal tissues is either inherited from parents or caused by de novo amputations. Many DCV occurs in Low Copy Repeat segments (LCR), which are also known as Segmental Duplication (SD) regions. In these regions, a sequence of 10-300kbp repeats itself multiple times and the replicate copies share at least 95% of their sequences. The DCV that encompasses a gene causes some alteration in the gene's production which may accounts for a significant portion of phenotypic variations [2]. Several studies presented evidences of high correlation of CNV with behavioral and developmental abnormalities such as cognitive impairment,

autism, mental retardation, and possibly psychiatric diseases [3]. Cohesive understanding of the DNA copy number and its production is the key to a better understanding of human diseases and phenotypes.

The importance of understanding the human genome is reflected in tens of thousands of studies being published every year to contribute to the human genome projects. The genomic hybridization experiment of new DNA samples is in high demand by numerous labs and clinics throughout the world. The copy number variation in these samples is detected using several algorithms and software packages. But these algorithms are not immaculate and they need a lot of improvement. Most of the algorithms fall short of achieving an acceptable amount of sensitivity or they generate large amounts of false calls. There is still an insistent demand for improvement in the field and this is our ultimate goal in the first part of this work.

In the second part of this thesis, we investigate patterns of the DNA copy number variations. The human genetic network is a quite complex system where hundreds, or even thousands, of DNA segments interact internally with each other directly or indirectly to control all body's functions. Each specific trait or phenotype is affected by several parts of the genome at the same time. Efficient clustering techniques after accurate detection of the DCN variation can provide biological signatures of the studied phenotype.

1.3 Contributions of this work and chapter descriptions

In chapter 2, we present four new algorithms to detect the variation of the DNA copy number under the conventional two-channel approach. The algorithms are: *Truncated Maximum Likelihood Test* TMLT, *Minimum Interval Score* MIS, *Uncovered Markov Model* UMM, and *Band-Pass Wavelet Transform* BPWT. We prove their superiority over 25 existing algorithms and software packages. Then, we discuss the reproducibility of the ROC curves from one experiment to another.

In chapter 3, we investigate three sources of non-biological bias in the DCN microarrays. We present three models: *Universal Threshold Adjustment UTA, GCNORM,* and *FLNORM* to remove the bias of the imperfect scanner, the GC content, and the fragment length, respectively. Next, we present a novel *Quantile-based Perfectly Isolated* model QPI of the distribution of the microarrays. And also we prove that the hybridization process is stationary and show the impact of this result on the analysis.

In chapter 4, we introduce the first single-channel approach for the analysis of the DNA microarrays. The approach detects and quantifies the variation of the DNA copy number using sensor networks approach. We expand the theory of the saddle-point approximation to cover the non-homogenous environments like DNA microarrays. We present *Log Lattice Lemma* (LLL) to maximize the performance of the non-uniform scalar quantizers. We prove that a quadratic quantizer followed by a moving average window is capable of analyzing the DCN arrays more accurately than any existing two-channel method.

In chapter 5, we present the *Segregation-Based Subspace Clustering* algorithm SBC to identify specific patterns of DCN variations. We will show the connection between the variations and autism and advance prostate cancer.

Chapter 2

DNA Copy Number Profiling Using Two-Channel Approaches

2.1 Genomic Hybridization Experiment

The genomic hybridization experiment is a powerful biological tool to measure the level of the DNA copy number in thousands or millions of genomic sites simultaneously. The experiment follows a standard approach that consists of sequential steps. The first step is to dismantle the DNA molecules into short fragments of length 0-2000bps using specific types of restriction enzymes such as NspI and StyI. After that, the short fragments are ligated to a very short sequence (~4bps) that can be recognized by the polymerase chain reactor PCR. The PCR amplifies the ligated fragments by producing thousands of identical copies of them to make the quantities of the DNA fragments readable or detectable. The amplified fragments are purified and denaturized using heat to separate the two DNA strands from each other. Only one of the two strands is taken into consideration during the experiment while the other strand is renounced. The single-stranded fragments, which are called targets, are dyed with a special fluorescence and finally hybridized to a fabricated chip.

The chip consists of made up probes that match partial sequences of the renounced strands to make the targeted strands bind to them. The targets from each specific site of the genome (ideally one target from each chromosome of each DNA molecule) are captured by thousands of identical complementary probes on the chip. The number of fabricated probes is much larger than the number of targets. When the hybridization process is completed, the chip is washed to remove the renounced fragments. And finally, the chip is scanned using a high-resolution scanner to generate an image whose intensities are equivalent to the amount of DNA fragments that have bound to each specific probe. Each probe on the chip is represented by a single pixel in the image. The intensity of that pixel represents the copy number level at a certain site of the genome. These intensities are the final product of the hybridization experiment and they form the signal that is used in the detection and pattern recognition analysis.

The probe design is a very tedious process and it depends on multiple factors. Since the binding between targets and probes is affinity-based, a target might bind to a probe other than its specific capturer if they match partially. This phenomenon is widely called cross-hybridization and it produces a large amount of error in the data. The main goal of the probe design is to reduce the similarity among targets to avoid the cross-hybridization. The experiment's resolution is equivalent to the total number of probes on the chip or to the average number of probes per a unit length of the genome. Consequently, the average length of the targets must be reduced to increase the resolution, and that endorses larger components of cross hybridization. There is a trade-off between the resolution and the error component. The number of probes that can be designed on one chip reaches up to 1.8 million probes in some platforms and that provides a good resolution to detect fine variations.

All available detection methods employ a conventional two-channel approach. The approach is based on a comparison between a test and a reference DNA samples and thus the process is called comparative. The intensity ratio of the test and the reference at each probe is analogous to the ratio of their copy number at the corresponding site on their genomes (R = test/reference). The ratio is greater than one if the test sample has

gained more copies than the reference, less than one if the test sample has lost one or more copies, and equal to one if they have the same copy number. The ratios are usually transformed to the $\log_2$ space where the duplication corresponds to positive values, the deletion corresponds to negative values, and the no-change status corresponds to zero values.

The next-generation sequencing is an emerging powerful method to read the whole sequence of a DNA molecule. It is by far the most precise and accurate tool of reading the genome's nucleotides but at a significantly higher price. The current price range of the next-generation sequencing methods is around $13,000 per sample (or $6,500 for one set of 23 chromosomes) while the cost is around $300 for genomic hybridization experiments [4].

2.2 Data Modeling

The variation of DCN usually spreads into segments covering several probes. Thus, all the included probes in the variant interval, theoretically, have the same intensity. Therefore, the comparative profile is usually modeled as a piecewise function consisting of constant sub-functions. The mean of every sub-function is unknown and it is equal to the copy number ratio of the test and the reference at that sequence. The locations of the transitions between the sub-functions are unknown as well.

If the experiment was ideal and the reference sample did not have any copy number variation, the mean of the sub-functions would be a discrete random variable with values equal to $\log_2(\chi/2)$ where χ is a non-negative integer. However, the cross-hybridization component does not have a zero mean, and the reference sample does not necessarily have only two copies everywhere in its genome. Therefore, the mean of the sub-functions of a noise-free comparative profile is a continuous random variable.

In real experiments, the $\log_2$ ratios are corrupted with substantial amounts of noise from many sources, and the variation is not even visually seen in the plots. Therefore, the goal of partitioning the genome into segments of the same DNA copy number is equivalent to partitioning the noisy $\log_2$ ratios into a piecewise function with unknown transition locations and amplitudes. The amplitude of the sub-functions in the $\log_2$ domain is 0 in normal cases, and a positive or negative real number in the variant regions. Higher amplitude corresponds to a higher variation. The noise is widely assumed to be additive white Gaussian.

$$Z[n] = F[n] + W[n] \text{ , } n = 0,1,2,\ldots,N-1 \qquad (2.1)$$

Where F[n] is the true piecewise function, Z[n] is the observed noisy signal, and W[n] is the additive white noise $N(\mu,\sigma^2)$. The challenge is to extract the noise-free signal F[n] from the observation Z[n]. F[n] consists of M successive segments, each segment has an unknown start, end, and mean.

Accurate identification of the break points between the sub-intervals is the most crucial step in the process. The remaining step is just to replace the raw intensity ratios in each segment by their arithmetic means.

2.3 Related Work

The study of detecting the alteration of DNA copy number has drawn a lot of attention in the last decade. The importance and the high resolution of the CGH arrays have attracted researchers to develop tens of algorithms to analyze the copy number microarrays [5-25]. The algorithms can be categorized into four famous approaches: Finite Impulse Response (FIR) filters, Hidden Markov Models (HMM), Neyman-Pearson theory tests, and Maximum Likelihood Estimators (MLE).

2.3.1 Finite Impulse Response Filters

Finite impulse response filter (FIR) is a general concept that comprises various models such as moving weighted average windows, random walk process, wavelet transforms, and others. The FIR filter is applied directly at the $\log_2$ ratios, Z[n]:

$$Z'[n] = Z[n] * h[n]$$

The uniformly weighted moving window is one of the earliest FIR filters used in the microarrays [26] because of its simplicity and time efficiency. The filter's coefficients can be uniform or non-uniform, and its order can be constant or variable. Sigma filter [5] is an example of uniform-coefficient variable-order FIR filters. The filter's order varies because it eliminates the observations that exceed local thresholds. Only the retaining observations are averaged. The filter in CGHRW [6] and SegN [7] is a step function which makes the process equivalent to the random walk model. The output of the filter is segmented based on local trends that identify the breakpoints.

The discrete wavelet transform [8-11] is a very popular application and it consists of a bank of FIR filters. The wavelet coefficients are computed as:

$$Wf(j,n) = \sum_{u=-\infty}^{\infty} \Psi_{j,n}[u].Z[u]$$

Where $\Psi_{j,n}[u] = 2^{j/2}\Psi([u-n]$. The term $\Psi_{j,n}$ is the Haar wavelet in the *Maximal Overlapping Discrete Wavelet Transform* MODWT [8] and it is the first derivative of Gaussian wavelet in GWT [9]. Another way of creating the filter bank is by the combination of multiple hierarchical levels of atomic filters like Haar as shown in figure 2.1. The *Discrete Wavelet Transform* (DWT) and the *Stationary Wavelet Transform* (SWT) [10] are hierarchical structures to decompose the signal into several frequency bands. The low pass filter L is $[1/\sqrt{2} \;\; 1/\sqrt{2}]$ while the high pass filter H is $[-1/\sqrt{2} \;\; 1/\sqrt{2}]$. The filters are the same at all levels of DWT while they are upsampled at each level of SWT. The wavelet coefficients are filtered using hard or soft threshold, then the coefficients are inversely transformed to re-construct the denoised signal.

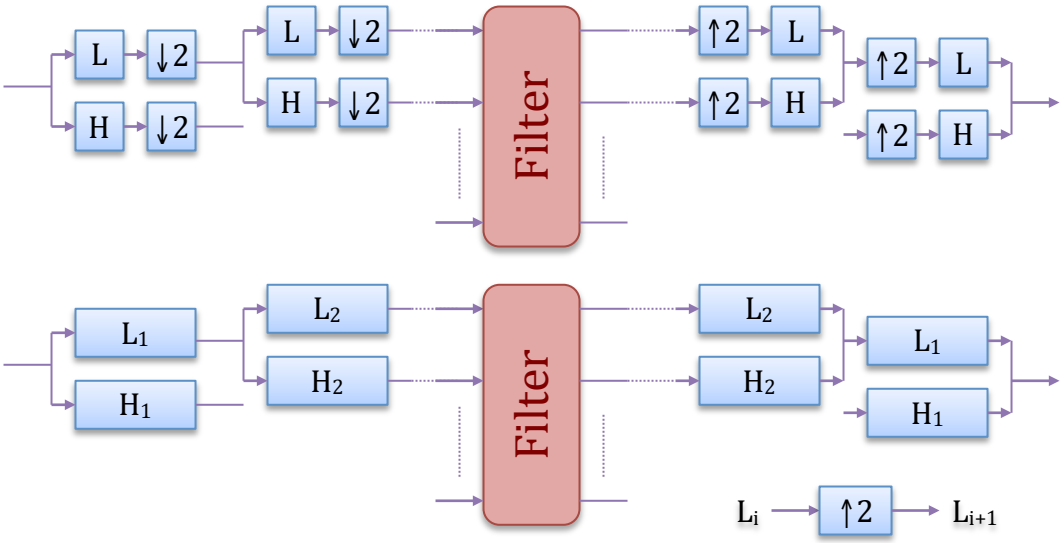


Figure 2.1: The filter banks of the discrete wavelet transform (top) and the stationary wavelet transform (bottom)

Other methods like WaveCD [11] employ adaptive thresholds to filter the wavelet coefficients. The thresholds are extracted from the coefficients at each level of the bank.

2.3.2 Hidden Markov Models:

Hidden Markov Models are governed by three factors: the hidden states, the transition matrix, and the initial state distribution [27]. The model assumes that the observations have only one order of dependency (Markov) and their underlying comparative copy number measurements are represented by the hidden states of the model [12-14]. The transition probability matrix rules the transition from one state to another, and all states are assumed to be connected. The transition matrix allocates most of its weight to remaining in the same state and it allocates small non-zero probabilities to transitioning to other states. The model can be seen as a clustering algorithm where the observations' order is preserved. The model employs Expectation-Maximization algorithms (EM) to construct the hidden states and the transition matrix that maximizes the likelihood between the model and the observations. The noise distribution is either assumed normal or extracted from the data.

The model was first introduced to the microarrays field by [28] and that model is embedded in the package FASST [12]. SMAP [13] is another hidden Markov model which takes into consideration the genomic distance between the targets. The dependency decays as the distance increases. Both FASST and SMAP assume that the number of states is known a priori. CGHRJA [14] takes into consideration the genomic distance and presents a higher level of sophistication by inferring the number of states from the observations.

2.3.3 Maximum Likelihood Estimators

The ultimate goal of the maximum likelihood estimators (MLE) is to estimate the true noise-free piecewise function that represents the comparative genomic profile. The bottleneck for these estimators is the heavy computational load since a dataset of size N observations can be segmented into 2^{N-1} different piecewise functions. The piecewise function that exhibits the maximum likelihood with the observations is selected. Several dynamic programs and clustering techniques were introduced to reduce the complexity of the solution [15-24]. If the noise is Gaussian, the maximum likelihood coincides with the least square error. Therefore, the solution X for the observations Y_i, i = 1,2,…,N is the one that minimizes the quantity:

$$\operatorname{argmin} \sum_{i=1}^{N} \|Y_i - X_i\|^2$$

Intuitively, the maximum likelihood occurs when the piecewise function is broken into N segments where each piece of the function consists of only one observation. In such a case, the likelihood is 1 and the error is zero, but the solution is meaningless. A stopping function must be employed to avoid overestimating the solution. CGH-segmentation [30] and CGHtrimmer [21] apply similar dynamic programs with various stopping functions: Akaike information criterion (AIC) [16], Bayesian Information Criterion (BIC) [17], Emillie Lebarbier [18], and Marc Lavielle [19]. The first three functions are merely penalty terms to penalize the likelihood for adding one more breakpoint. Therefore, the previous quantity becomes:

$$\text{argmin}\left\{\sum_{i=1}^{N}\|Y_i - X_i\|^2 + \text{Penalty term}\right\} \qquad (2.2)$$

The penalty term for fitting the data into K segments is equal to 2K in AIC, Klog(n) in BIC, and 2.6log(N/K)+2 in Emillie. The fourth stopping criterion is adopted from the dynamic program itself to make the segmentation process stop when the improvement of the likelihood is not significant. The penalty term of CGHtrimmer is similar to AIC. BCP [22] and cghFLasso [20] are other dynamic programs based on Barry and Hartigan model [31] and SQOPT algorithm [32], respectively.

The process of the clustering algorithms is similar to the dynamic programs. The only difference is that the clustering algorithms are bottom-up approaches as opposed to the up-bottom approaches in the dynamic programs. The dynamic programs start with one segment covering the whole observations and they break it progressively into smaller segments. The clustering algorithms start by assigning every observation in a separate segment and then they combine the small segments hierarchically based on their similarity until the observations are combined into one segment. Several techniques are used to stop the clustering process. CLAC [23] employs a universal threshold where only the clusters above that threshold are considered. Vega [24] adopts the exact opposite stopping criterion of Marc. Iteratively, the algorithm combines the two segments whose impact on the likelihood is the least. It keeps combining more segments, as long as the deterioration of the likelihood is not significant.

2.3.4 Neyman-Pearson Theory

The Neyman-Pearson theory is a powerful tool for the two-hypothesis tests to dictate the rejection of the null hypothesis. The rejection is determined based on a score given to the test. The test becomes the uniformly most powerful test if the test is one-sided, i.e. the detection of the duplication is performed separately from the detection of the deletion.

If the noise process is independent and identically distributed (i.i.d) Gaussian with zero mean and unit variance, then the score of an interval of M observations is equal to their sum divided on the square root of M: $\text{score} = \sum_{i=1}^{i=M} Y_i/\sqrt{M}$. The interval score is directly proportional to the level and the length of the variation of the copy number. Circular Binary Segmentation (CBS) [29] is one of the most popular and widely used algorithms in the analysis. It is embedded in several packages like DNAcopy [15] and RANK [12]. The algorithm seeks the interval whose score is the highest, and then it looks for the interval whose score is second highest and so on. N^2 iterations are required to detect each interval where N is the number of observations. Several solutions were suggested in [33] to reduce the computational load.

2.4 Our Contribution

The continuing increase of the technology of the CGH arrays leads to a constant increase in the data size and to more necessity for simpler but efficient algorithms. We present four novel methods to analyze the DCN microarrays: *Band-Pass Wavelet Transform* (BPWT), *Uncovered Markov Model* (UMM), *Truncated Maximum Likelihood Test* (TMLT), and *Minimum Interval Score* (MIS). Each method belongs to one of the main four categories.

2.4.1 Band-Pass Wavelet Transform BPWT

The structure of the filter bank in figure 2.1 is equivalent to an orthonormal matrix *W* of size 2^L where L is the number of decomposition levels. The elements of the matrix are $\pm 1/2^{L/2}$. The wavelet coefficients are computed through a regular convolution process:

$$Wf(n,j) = \sum_{u=-\infty}^{\infty} W[j].Z[n-u]$$

The wavelet transform is widely used in data compression, image processing, filtering audio signals, and several other signal processing applications. The treatment of the wavelet coefficients is modified based on the problem that is being solved. In the analysis of the comparative DCN, the piecewise constant function is featured with very low frequency components. Therefore, the bases of the matrix W should be given different weight in the analysis, where the low-frequency bases are given more weight than other bases. We chose to give weight 1 to the low frequency bases and zero weight to the others. We define the low frequency base as the base whose elements' sign changes at most one time. The column W_j is considered a low-frequency base if it satisfies the bound:

$$\sum_{i=1}^{2^L-1} \left| \; W[i+1,j] - W[i,j] \; \right| \leq \sqrt{2^{2-L}} \qquad (2.3)$$

The microarray data sets are subject to great amounts of noise, and the signal-to-noise ratio is relatively small. Therefore, the wavelet coefficients of non-low frequency features are definitely generated by noise and therefore, they must be eliminated.

$$Wf(n,j) = \sum_{u=-\infty}^{\infty} \mathrm{C}_j \mathrm{W[j]}.Z[n-u] \qquad (2.4)$$

$$\mathrm{C}_j = \begin{cases} 1, & if\ W_j\ satisfies\ (2.3) \\ 0, & Otherwise \end{cases}$$

A universal hard threshold is applied at the wavelet coefficients:

$$Wf(n,j) = \begin{cases} Wf(n,j), & Wf(n,j) \geq 2median\ (|Wf(n,j)|) \\ 0, & Otherwise \end{cases} \qquad (2.5)$$

And the piecewise function is re-constructed as:

$$\hat{Z}[n] = \sum_{j=1}^{2^L} \sum_{u=-\infty}^{\infty} \mathrm{W[j]}.Wf(n-u,j)$$

2.4.2 Uncovered Markov Model

The hidden Markov Model (HMM) can be regarded as an unsupervised clustering algorithm. The clusters' centroids are analogous to the hidden states while the order of the observations is preserved. The essential fault of applying this model into the DCN microarrays is that there are no real "states" in the data, especially in the low copy repeat (LCR) sequences. The ideal comparative copy numbers in the LCR sequences take rational values (C_{test}/C_{ref}) where C_{test} and C_{ref} are non-negative integers. That means the number of possible states is relatively larger than the number of states that can be analyzed practically by the available algorithms. Therefor, the observations are forced to fit into an under-estimated model and consequently, the small copy number variation will not be detected.

In our uncovered Markov model (UMM), the *actual* number of states is totally ignored. All duplication states are substituted by one state representing the minimum gain and all deletion states are substituted by one state representing the minimum loss. The mean vector $\boldsymbol{u} = (-u, 0, u)^T$. The initial state distribution is not critical to the analysis and we chose it to be $\boldsymbol{\pi} = (1/3\,, 1/3\,, 1/3)$. The state transition probability matrix A is:

$$\boldsymbol{A} = \begin{bmatrix} 1-2\epsilon & \epsilon & \epsilon \\ \epsilon & 1-2\epsilon & \epsilon \\ \epsilon & \epsilon & 1-2\epsilon \end{bmatrix}$$

And the emission distribution B has independent unit variance Gaussian distributions:

$$\boldsymbol{B} = \begin{bmatrix} N(-u,1) \\ N(0,1) \\ N(u,1) \end{bmatrix}$$

The expectation-maximization algorithm is eliminated since the states are already uncovered, and that is a considerable reduction in the complexity. The model that exhibits the maximum likelihood with the observations under the parameters **A**, **B**, and $\boldsymbol{\pi}$ is selected to be the solution. The value of the likelihood is measured using the Forward-Backward procedure [27].

2.4.3 Truncated Likelihood Ratio Test (TLRT):

The estimator that is used in the dynamic programs and the clustering algorithms is significantly redundant. If we look at the noise-free piecewise function, we find that almost one half of its components are already known and they have zero amplitude representing the normal state which fits with the baseline. The quadratic complexity of estimating all pieces of the solution can be reduced to a linear complexity by focusing only at the encompassed variant pieces. Therefore, the estimating process converts to a detecting process without jeopardizing the performance.

The role of the first term of equation 2.2 can be achieved using a moving likelihood ratio test against one-sided hypothesis. The role of the second term, which is responsible for reducing the effect of the outliers, can be incorporated into the likelihood ratio test by truncating its extreme values. Assuming that the noise process is i.i.d Gaussian with zero mean and σ^2 variance, the likelihood test of the hypothesis of an existing duplication H_1: $N(u_1, \sigma_1^2)$ of size M is:

$$\Lambda(Y) = \prod_{i=1}^{M} \frac{L(Y_i/H_1)}{L(Y_i/H_0)} = \left(\frac{\sigma_0}{\sigma_1}\right)^N \prod_{i=1}^{M} \exp\left\{\frac{(Y_i-u_0)^2}{2\sigma_0^2} - \frac{(Y_i-u_1)^2}{2\sigma_1^2}\right\} \underset{H_1}{\overset{H_0}{\lessgtr}} \frac{\pi_0}{\pi_1} \qquad (2.6)$$

The left hand side of (2.6) should be truncated to eliminate the effect of the outliers. If we define the operator $\lceil Y \rceil$ and $\lfloor Y \rfloor$ as:

$$\lceil Y \rceil = \begin{cases} Y, & Y < \tau_{upper} \\ \tau_{upper}, & Y \geq \tau_{upper} \end{cases} \quad \text{and} \quad \lfloor Y \rfloor = \begin{cases} Y, & Y > \tau_{lower} \\ \tau_{lower}, & Y \leq \tau_{lower} \end{cases}$$

Then the likelihood ratio test becomes:

$$\ell(Y) = \prod_{i=1}^{M} \left\lceil \exp\left\{\frac{(Y_i-u_0)^2}{2\sigma_0^2} - \frac{(Y_i-u_1)^2}{2\sigma_1^2}\right\} \right\rfloor \underset{H_1}{\overset{H_0}{\lessgtr}} \frac{\pi_0}{\pi_1}\left(\frac{\sigma_1}{\sigma_0}\right)^N \qquad (2.7)$$

If the noise is homogenous ($\sigma_0 = \sigma_1$), $u_0 = 0$, and $\pi_0 = \pi_1 = 0.5$, then the log likelihood ratio test of a duplication event is simplified to:

$$\mathrm{LLR(Y)_{dup}} = \sum_{i=1}^{M} \left[Y_i - \frac{u_1}{2} \right] \underset{H_1}{\overset{H_0}{\lessgtr}} 0 \qquad (2.8)$$

The test for the deletion hypothesis H_1: $N(-u_1, \sigma_1^2)$ is similar to (2.8) with a reversed relational expression as:

$$\mathrm{LLR(Y)_{del}} = \sum_{i=1}^{M} \left[Y_i + \frac{u_1}{2} \right] \underset{H_0}{\overset{H_1}{\lessgtr}} 0 \qquad (2.9)$$

This test is sufficient to identify the segmental variations more accurately than the dynamic programs and the clustering algorithms. And yet, it is nothing but a moving summation window of size M applied on the observations after truncating the ones that are significantly far from $u_1/2$. Sigma filter follows a similar approach, but it employs local thresholds instead of a universal one, and it eliminates the outliers instead of truncating them.

The algorithm SW is also based on the principle of truncated likelihood-ratio-test but with varying-sized window. It eliminates the observations whose absolute values are above a universal threshold and assigns zero values to the observations whose absolute values are under another universal threshold. The main misstep of this algorithm is that it permits using very short windows and that generates large amounts of false alarms in environments with low SNR.

The choice of M is controlled by two parameters: the false positive rate (false alarm P_f) and the signal-to-noise ratio (SNR) where the SNR is defined as $20\log_{10}(\mu/\sigma)$ in dB. Considering the independent and homogenous Gaussian noise, the relationship that governs the three parameters: M, P_f, and SNR ($\mu_1/2\sigma$) is:

$$P_f = \int_{u_1/2\sigma}^{\infty} \frac{1}{\sqrt[M]{2\pi\sigma^2}} e^{-\left(\frac{Mx^2}{2\sigma^2}\right)} dx \qquad (2.10)$$

The relationship is illustrated in figure 2.2 for multiple values of P_f.

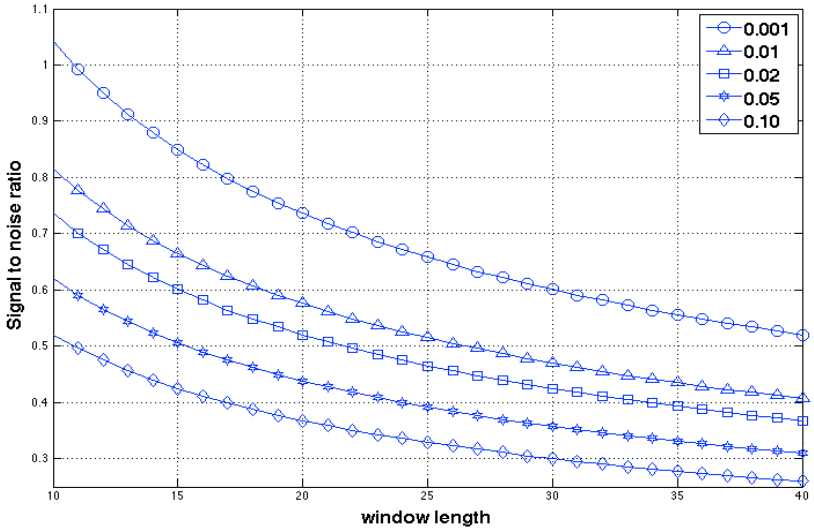


Figure 2.2: SNR (M/2σ) versus window size for multiple levels of false alarm

2.4.4 Minimum Interval Score:

As we mentioned earlier, the Most Powerful Test MPT, is performed by assigning a score to an interval of size M; score $= \sum_{i=1}^{i=M} Y_i/\sqrt{M}$. The maximum score coincides with the minimum false alarm if the noise distribution is Gaussian. MPT is uniformly the most powerful test if one of two requirements is met: either the length of the variation is accurately known, or if all observations in the window are drawn from the same distribution (H_1 or H_0). If none of the two requirements is met, which is the case in DCN microarrays analysis, then the results of the UMPT might be misleading even in noise-free environments. We present a simple example to demonstrate that fact.

Assume F is a piecewise function consisting of 2K+1 segments of different length F_1, F_2, …, F_{2K+1}. The magnitude of odd-numbered segments F_1, F_3, …, F_{2K+1} is μ (duplication) and the magnitude of even-numbered segments F_2, F_4, …, F_{2K} is zero (normal status). Also assume that F_1 is relatively longer than the other segments which implies that its score is larger. Under these assumptions, it is straightforward to conclude that the score of all the pieces of the piecewise function F (including the duplicated and the normal segments) is equal to:

$$\text{Score}(F) = \text{score}(F_1)\sqrt{\frac{1+\frac{O}{A+O}+\frac{O^2}{A(A+O)}}{1+\frac{E}{A+O}}}$$

Where O is the total length of the odd segments: $O = |F_1| + |F_3| + \cdots + |F_{2K+1}|$, E is the total length of the even segments: $E = |F_2| + |F_4| + \cdots + |F_{2K}|$, and A is the length of F_1. This result states that, if segments F_3 is longer than F_2:

If $|F_3| > |F_2|$

$\Rightarrow$ $\text{Score}(F_1 \cup F_2 \cup F_3) > \text{Score}(F_1)$

then if $|F_5| > |F_4|$

$\Rightarrow$ $\text{Score}(F_1 \cup F_2 \cup F_3 \cup F_4 \cup F_5) > \text{Score}(F_1)$

That means, for any K+1 duplication segments separated by K normal segments: if the total length of the K normal segments is shorter than the total length of the K+1 duplication segments, then the score of the 2K+1 segments is the highest. That means all the normal segments in the middle will be called copy number variant. That is the reason why CBS algorithm has a great tendency to combine close segments of duplication or deletion into one large segment, which generates a large amount of false alarms. We present a numerical example to demonstrate this result.

Consider the piecewise function $F = \{F_1 U F_2 U F_3 U F_4 U F_5\}$. $E[F_1]=E[F_3]=E[F_5]=10$. And $E[F_2]=E[F_4]=0$. The standard deviation = 0.25 for all pieces. $|F_1|=100$, $|F_2|=50$, $|F_3|=60$, $|F_4|=60$, and $|F_5|=40$. The observations are classified in distinct clusters as shown in figure 2.3 and it is easy to identify the breakpoints. However, the CBS algorithm combines the 5 segments in one large segment. The reason is that the score of the first segment, which is the highest segment, is 99.9 but the score of the five segments together is 113 in this example.

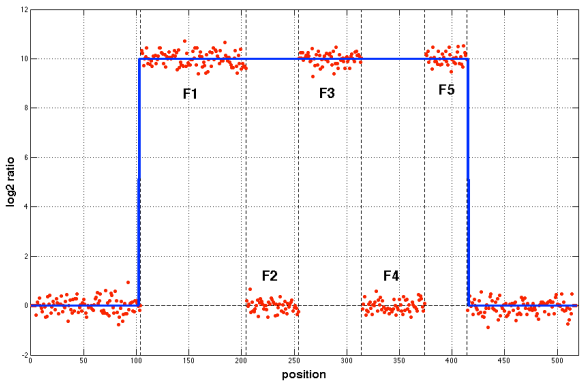


Figure 2.3: Numerical result of CBS (solid line) applied at the observation (dots)

Since the maximum interval score does not necessarily lead to the interval of the maximum variation, it is intuitive to replace the search for the maximum score by the search for all intervals whose scores exceed a universal threshold. This approach was mentioned briefly in [33] but has never been explored. Because the hypothesis is one-sided, the test is applied twice, one to detect the duplication and one to detect the deletion.

The size of the one-sided UMPT is measured by its specificity, $\Phi(\eta) = \int_{-\infty}^{\lambda} \varphi(x)dx$ where $\varphi(x)$ is a zero-mean unit-variance normal distribution. The outliers are removed by truncating the distant observation as we did in TLRT. If Y is an interval of width M, the minimum interval score test for the duplication hypothesis H_1: $N(u_1, \sigma_1^2)$ is:

$$\mathrm{MIS(Y)_{dup}} = \frac{1}{\sqrt{M}}\sum_{i=1}^{M}[Y_i] \underset{H_1}{\overset{H_0}{\lessgtr}} \eta \qquad (2.11)$$

And the test for the deletion hypothesis H_1: $N(-u_1, \sigma_1^2)$ is:

$$\mathrm{MIS(Y)_{del}} = \frac{1}{\sqrt{M}}\sum_{i=1}^{M}[Y_i] \underset{H_0}{\overset{H_1}{\lessgtr}} -\eta \qquad (2.12)$$

We will show that this test achieves the most powerful performance in our comprehensive comparison in the next section.

2.5 Comprehensive Comparison Using Real-World Data

The performance of each algorithm is greatly controlled by several tuning parameters. It is common for different studies to test the same data using the same algorithms but report significantly different results. That raises the importance of conducting a comprehensive and accurate comparison of the available algorithms to assess their performance under wide ranged parameters.

Several comparative analyses have been published to compare multiple algorithms [34-42]. The biggest comparison is the one made by Lai et al [35], which compared 11 algorithms and generated their receiver operating characteristic curves (ROC) using simulated data. None of the comparisons considered the detection of the low variations where the segment mean-to-noise ratio is less than one (SNR < 0dB).

Here we present a comprehensive comparative analysis of 29 algorithms using real world data. The algorithms are: Sigma [5], RANDOMwalk [6], SegN [7], MODWT [8], GWT [9], SWT and DWT [10], and WaveCD [11] from the FIR filters category. FASST[12], SMAP [13], and RJACGH [14] from the MMs category. SegMNT [15], AIC [16], BIC [17], Emillie [18], Marc [19], Flasso [20], CGHtrimmer [21], bcp [22], CLAC [23], Vega [24], and SW [25] from the category of the maximum likelihood estimators. And DNAcopy [15] and RANK [12] from the Neyman-Pearson category. We also included our algorithms BPWT, UMM, TLRT, and MIS.

2.5.1 Finely Tiled Arrays:

We got the data from [42]. Seven comparative arrays were designed by Roche-NimbleGen at a resolution of 1probe/120pb at segmental duplication regions and 1probe/200pb in the unique sequences. The arrays cover five genomic intervals from five chromosomes: chr7: 61058424-82000033, chr10: 77000071-91999959, chr15: 18260026-34999973, chr17: 12000112-22187066, and chr22: 14430001-26000041.

Chromosome	Center position	Source	Status
7	69835342	[43]	Gain
7	70059379	[43]	Gain
7	71910357	[43]	Loss
10	88959840	[45]	Loss
15	18841527	[45]	Gain
15	19153955	[45]	Gain
15	19191228	[43]	Gain
15	19911312	[43]	Gain
15	20231763	[43]	Gain
15	21609973	[43]	Gain
15	24990280	[43]	Gain
17	18306691	[1]	Gain
22	14529515	[45]	Loss
22	15238042	[43]	Gain
22	16438723	[43]	Gain
22	18839270	[1]	Loss
22	18966258	[1]	Loss
22	19227578	[1]	Gain
22	20783786	[43]	Gain
22	21002600	[43]	Gain
22	21027437	[45]	Gain
22	21291645	[43]	Gain
22	21344609	[43]	Gain
22	22684672	[1]	Gain
22	22715307	[1]	Gain

Table 2.1: QPCR sites for NA10851 versus NA15510

The total length of the five intervals is almost 75Mbps (almost 2.5% of the whole genome) with about 25% of their length covered by segmental duplication sequences. The total number of probes is 384,432 in each array. The authors in [43] conducted 4 dye-swapped experiments using DNA samples of two HapMap subjects: NA15510 and NA10851. In two experiments, they tested NA15510 versus NA10851 and in the other

two they tested NA10851 versus NA15510. That means the duplication in two comparative arrays appear as deletion in the other two arrays and vise versa. They also conducted three experiments to assess the rate of the false alarm. In each experiment, two DNA samples from the same individual are compared. Since the two samples are theoretically identical, any declared variation between them is merely a false alarm.

2.5.2 QPCR Test:

Various sites of NA10851 versus NA15510 were tested using the quantitative polymerase chain reaction test QPCR, and the results are reported in [1,43,45]. The QPCR is a highly reliable test to confirm the variation of DNA copy number if it truly exists, and it is widely used to evaluate the performance of detection. We selected 25 genomic variant sites confirmed by the QPCR to be included in our comparison. Since we have 4 arrays, the total number of confirmed sites is 25×4=100. The QPCR sites, sources, and statuses are presented in table 2.1.

2.5.3 Sensitivity Versus False Alarm:

For any algorithm, the sensitivity is measured as the percentage of the QPCR sites that are detected. The false alarm is measured as the total length of the detected variation in the three self-self arrays divided by their total length. The sensitivity and the false alarm rate are used to generate the receiver-operating characteristic curves (ROC) for the 29 algorithms.

We chose two measurements to evaluate the performance of the ROC curves: the area under the curve and the residual. The area under the curve (AUC) is the total area under the curve in the interval from P_f=0 to P_f=1. Its maximum value is 1 (perfect detection) and its minimum value is 0.5 (the no-discrimination line). An algorithm is considered good only if its AUC is equal to 0.9 or above. We define the residual as the total area above the curve in the interval [0, 0.1] multiplied by 10:

$$Residual = 1 - 10 \times \sum_{P_f=0}^{P_f=0.1} Sensitivity(P_f)$$

The residual also ranges from 0 to 1. It is more accurate to evaluate the performance only at low false alarm rates since the performance at high rates of false alarm is not considered. More efficient algorithms generate values of residual closer to zero.

2.5.4 Data Modeling

Two assumptions are widely considered in the analysis: the independency of the observations and the normality of the distribution. The first assumption can be verified by inspecting the autocorrelation of the self-self arrays as illustrated in figure 2.4. The autocorrelation is 1 at x=0 and almost zero everywhere else which confirms that the observations are independent. The second assumption can be verified by comparing the quantiles of the self-self experiments with the quantiles of a normal distribution. The Q-Q plot is illustrated in figure 2.5 and the straight line confirms that the two distributions are similar.

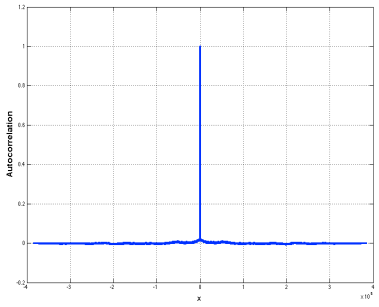


Figure 2.4: The autocorrelation of self-self experiment # 1.

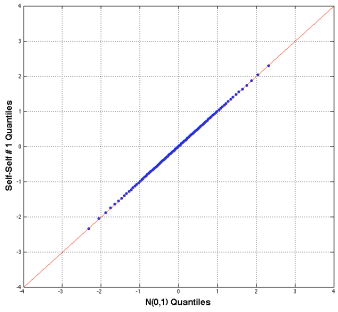


Figure 2.5: Q-Q plots of a self-self array versus a normal distribution

The variance of the $\log_2$ ratios is different from one experiment to another and therefore, the samples need to be normalized to have identical variances. Since the outliers have a great impact on calculating the variance, we scaled all our arrays to have a median absolute deviation equal to 0.6745. $MAD(x) = Median(|x - median(x)|)$. MAD = 0.6745 for unit-variance Gaussian distributions. This measurement is more robust to the outliers than the standard deviation. We also assume that the variance is the same in the duplication, deletion, and normal states, $\sigma_0=\sigma_1=1$.

2.5.5 Results and Discussion

We evaluated the sensitivity and the false alarm of the previously mentioned algorithms to generate the receiver operation characteristic (ROC) curves. The performance is relatively poor in SegN, GLAD, SW, MODWT, GWT, DWT, WaveCD, bcp, SegMNT, Emillie, BIC, and AIC as illustrated in figure 2.6.

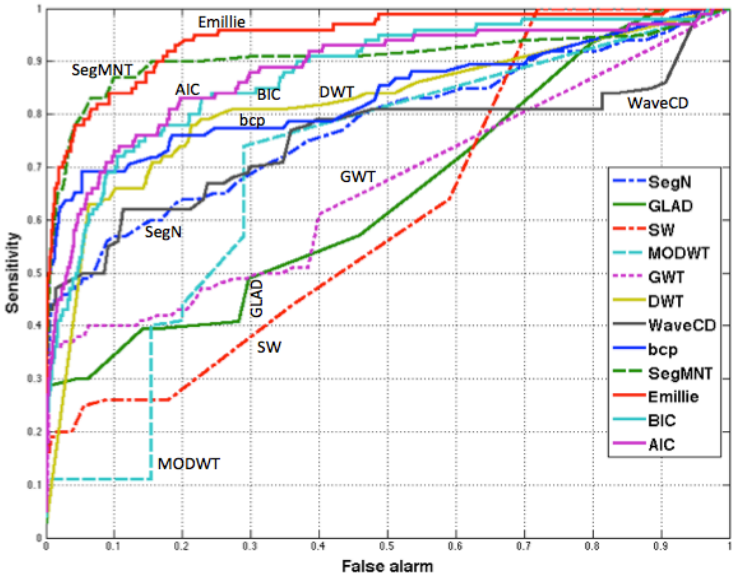


Figure 2.6: ROC curves of several poorly performing algorithms

SWT and DWT are almost identical except that SWT is $2^J N$ times redundant where J is the decomposition level and N is the data size. In [10], J is recommended to be equal to $\log_2(N)-4$ and it is justified because the noise power gets reduced to 2^{-J}. However, this is not true at low SNR because after a few decomposition levels, the wavelet coefficients of the true signal are drowned under the coefficients of the noise. In our experiment, the optimal choice of J is 4 for DWT and BPWT, and 5 for CWT. The results in figure 2.7 show that the redundant representation SWT is much better in its performance than the non-redundant DWT. It also shows that the performance of our algorithm BPWT enjoys some advantage over SWT at low false alarms. The two curves merge together when the false alarm > 0.1.

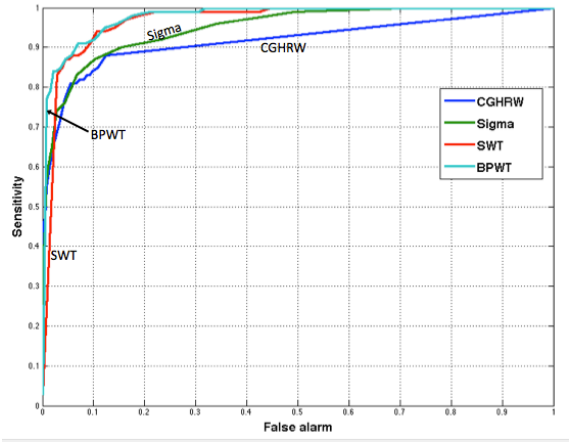


Figure 2.7: ROC curves of 4 FIR filters

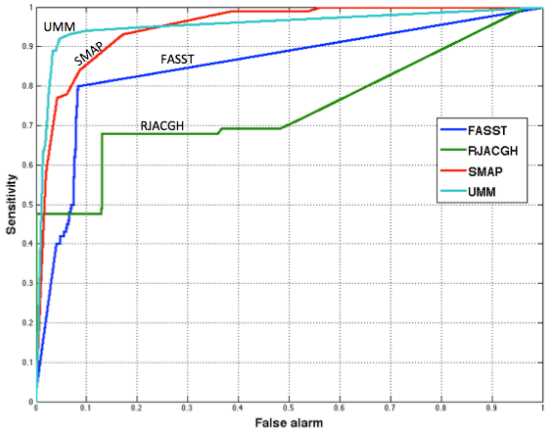


Figure 2.8: ROC curves of 4 Markov Models

Figure 2.8 shows that SMAP, FASST, and CGHRJA performed modestly even though SMAP's AUC is equal to 0.92! This is a practical example to show the benefit of using the residual along with the AUC. The high value of SMAP's AUC is concentrated mostly at high rates of false alarm covering the interval [0.2, 1]. However, the interest of the researchers is limited to narrow intervals on the left side of the ROC curve. The residual has large values in SMAP, FASST, and CGHRJA. The reason why all HMM algorithms perform modestly is that the solutions were forced to fit with under-estimated models as we explained earlier. The performance improved significantly when this restriction was eliminated in our algorithm UMM.

For the MLE algorithms, the performance of Vega is appealing. It is worth mentioning that the optimal parameter for Vega in our analysis is different than what the authors suggested in [24]. They assigned ±0.2 threshold to declare the gain and the loss but we found the threshold to be more accurate at ±0.5. The latter threshold generates 6.5% of false alarm compared to 33.8% using ±0.5.

In the dynamic programs, the largest AUC was achieved by CGHtrimmer with $\lambda = 1$ and the least residual was achieved by cghFLasso. The reduction in the waiting time is highly remarkable in CGHtrimmer compared to Picard's program, mainly due to their novel approach of building the NxN auxiliary matrix. The penalty function of AIC, bcp, cghFLasso are very cheap and therefore, the number of the detected segments is too large. On the other hand the algorithms SegMNT, Marc, and Emillie are conservative in breaking new segments and that leads to underestimated solutions. Approximately, the average number of segments per array are 240, 460, 560, 2600, 9800, 15000, and 22500 for SegMNT, Emillie, Marc, CGHtrimmer, BIC, cghFLasso, bcp, and AIC, respectively. Almost half of the segments in the last three algorithms are single outliers. Figure 2.9 proves that our algorithm TLRT is in agreement with Vega and Flasso at low false rate levels (P_f<0.05). TLRT surpasses them in the rest of the interval even though the range is not considered in any practical analysis.

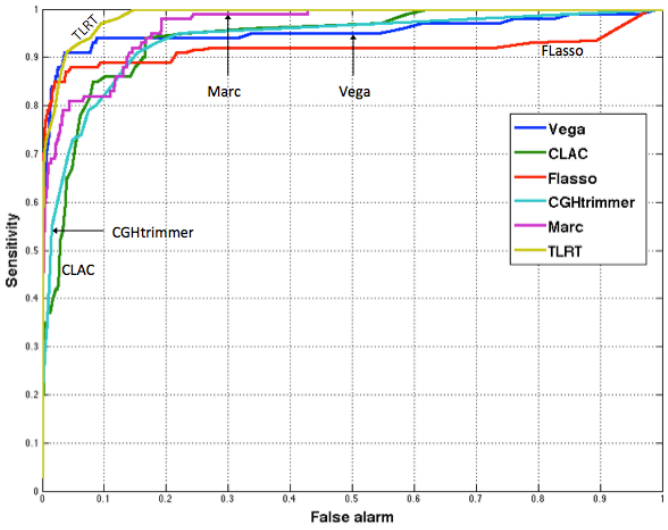


Figure 2.9: ROC curves of 6 MLE models

In the likelihood ratio test algorithms, the impact of replacing the varying sized window with a fixed sized one is remarkable. Regardless of the reduction in the waiting time, the area under the curve surged from 0.62 in SW to 0.98 in the TLR test, and the residual dropped from 0.77 to 0.12.

Figure 2.10 illustrates the ROC curves of Neyman-Pearson algorithms. It shows that, out of the 28 algorithms, our Minimum Interval Score (MIS) algorithm yielded the best ROC curve with the highest AUC and the lowest residual. It also shows that DNAcopy performed outstandingly better than RANK even though they implement the same CBS algorithm. This result highlights the sensitivity of Neyman-Pearson theory to the outliers. The improvement in the performance of DNAcopy is due to a preprocessing step to reduce the effect of the outliers using a median window. Still, it generated a high amount of false alarm because of its tendency to combine the adjacent small segments into one large segment as we proved earlier.

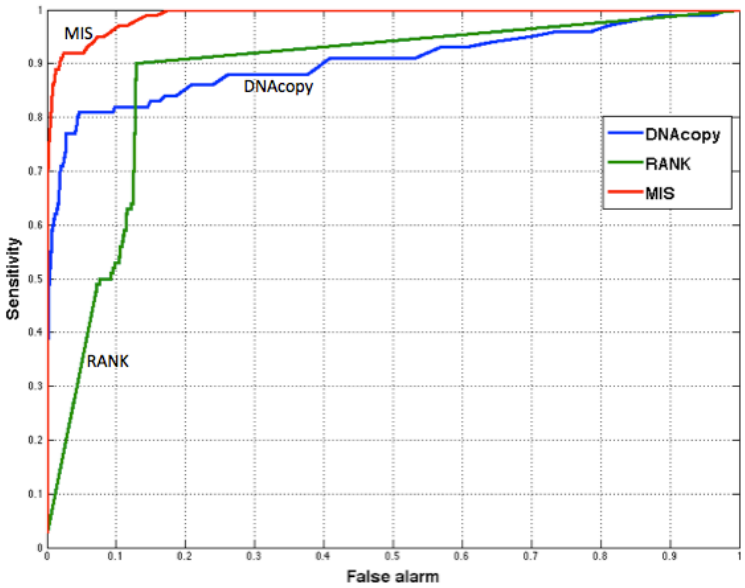


Figure 2.10: ROC curves of 3 Neyman-Pearson models

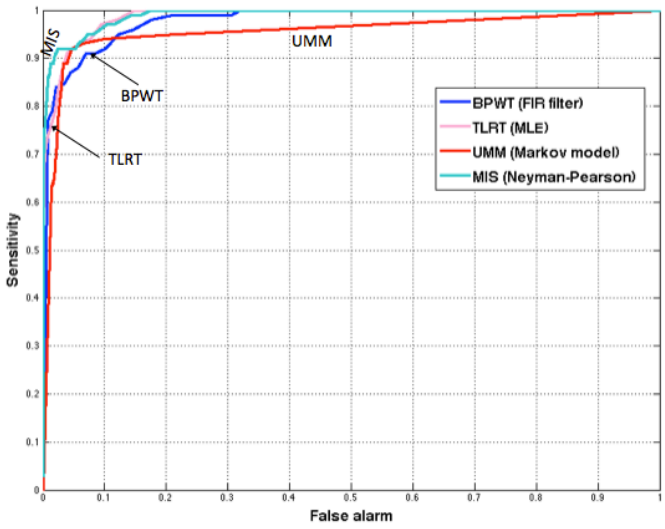


Figure 2.11: ROC curves of our BPWT, TLRT, UMM, and MIS

Figure 2.11 illustrates a comparison of our four algorithms. MIS and TLRT are in agreement in the interval [0.06, 1] but the false alarm of MIS is less than TLRT in the interval [0, 0.06]. A summary of the AUC and the residual of each algorithm is presented in figure 2.12. Based on the performance, our four algorithms are ranked the first, the second, the third, and the fifth in the list.

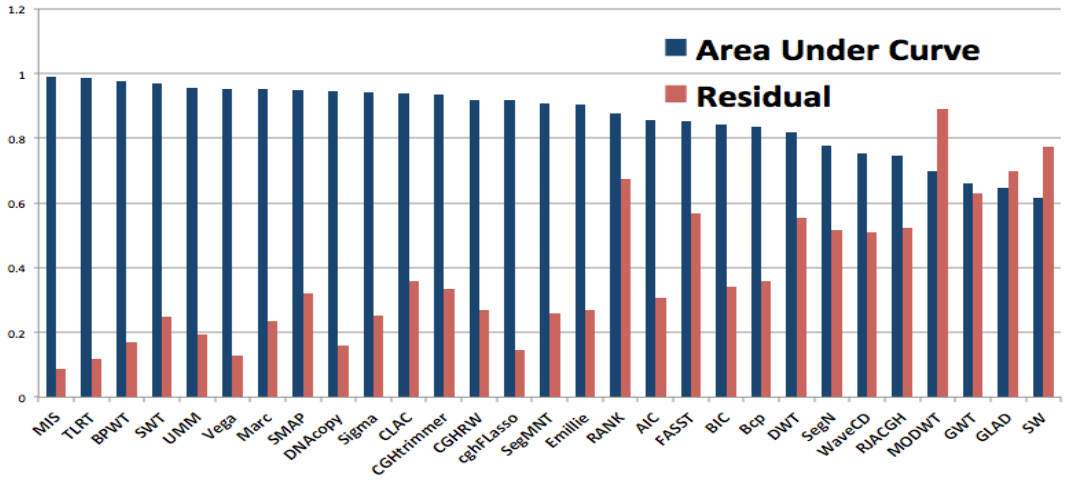


Figure 2.12: AUC (long bars) and residual (short bars) for various algorithms

2.6 Complexity of BPWT, UMM, TLRT, and MIS

The novel algorithms we presented are not only better in their detection performance, but also they are much more efficient in the computational load. If the data contain N observations and the filter bank has L decomposition levels, then the SWT requires $N(2^L-1)2^{L+1}$ additions and $N2^{2L+1}$ multiplications. Our algorithm BPWT requires only $4N(2^L-1)$ additions and $N2^{L+2}$.

Considering k states, HMMs require $2k(k-1)N$ additions and $2(k^2+k)N$ multiplications for each iteration of the EM algorithm. The algorithm requires hundreds of iterations to converge to the solution. This algorithm is not needed anymore in our algorithm UMM.

The MLE algorithms require N(N-1) additions and N(N-1) multiplications to built the auxiliary matrix. Using likelihood ratio test of size M, TLRT requires only 2NM additions with zero multiplications.

The CBS algorithm requires N(N-1) additions and N(N-1) multiplications. Using a hypothesis test of size M, MIS requires 2NM additions and 2N multiplications.

In our experiment, we had 385,000 probes divided into 5 intervals. Considering N_{avg}=77,000, M=25, and k=6 with 1000 iterations as the result provided by CGHRJA, the total numbers of required additions and multiplications are shown in table 2.2.

Existing algorithms		
Algorithm	Additions $\times 10^6$	Multiplications $\times 10^6$
SWT	36.96	39.42
CGHRJA	4620	6468
Vega	5929	5929
CBS	5929	5929

Our contribution		
Algorithm	Additions $\times 10^6$	Multiplications $\times 10^6$
BPWT	4.62	4.93
UMM	0.92	1.85
TLRT	3.85	0
MIS	3.85	0.154

Table 2.2: Additions and multiplications required by several algorithms.

2.7 Reproducibility of ROC curves

We conducted another experiment to test the similarity among ROC curves when they are generated by different datasets. Each point on the ROC curve is created by a specific tuning parameter and we want to test if using the same tuning parameter for the same algorithm at another dataset yields the same sensitivity and false alarm probability or not.

We used publicly available data from http://www.ncbi.nlm.nih.gov/geo/ with GEO accession GSE28111 [93]. The data contain 36 test-reference arrays of NA15510 versus NA10581 and 30 self-self arrays. Each array covers the whole genome. We used a list of 50 DCN variant sites confirmed by the QPCR [1]. The total 1800 (36x50) confirmed sites of the test-reference arrays are used to measure the sensitivity while the 30 self-self arrays are used to measure the probability of the false alarm. The ROC curves of our algorithms: MIS, TLRT, UMM, and BPWT are shown in figure 2.13.

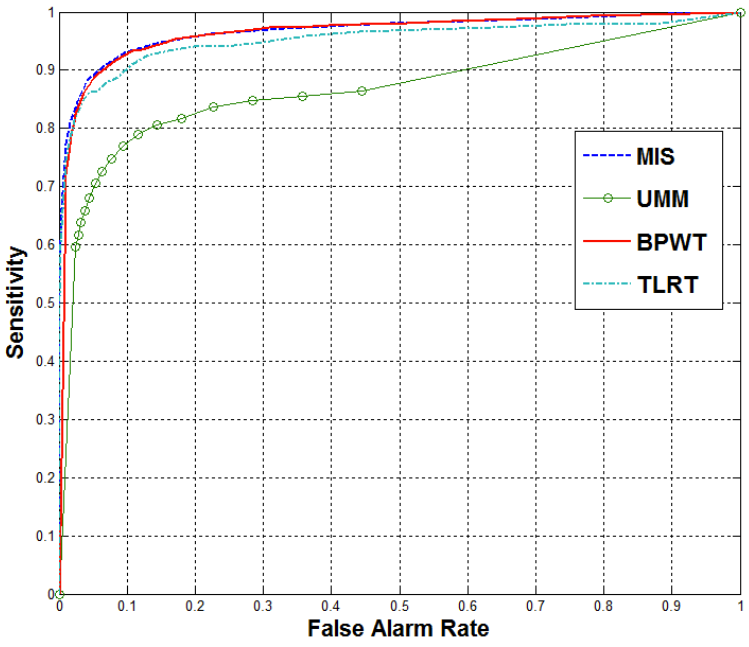


Figure 2.13: ROC curves for MIS, UMM, BPWT, and TLRT

To test the hypothesis, we applied a cross-validation approach on the data bank. In each trial, we compared the sensitivity from one single array with the sensitivity from all other arrays. The comparison is conducted by applying the same tuning parameter at the training arrays and the tested array to compare the values of sensitivity that they achieve. The same approach was applied to compare the false alarm rate between the tested and the training data. The tuning parameter's space is wide enough to permit the sensitivity or the false alarm probabilities to reach its limit of 0 or 1.

The sensitivity exhibits a reasonable stability in the cross-validated arrays in all four algorithms. The sensitivity-sensitivity plots between the tested and the training data generated by using the identical tuning parameters for MIS are shown in figure 2.14. The plots are similar in TLRT, UMM, and BPWT.

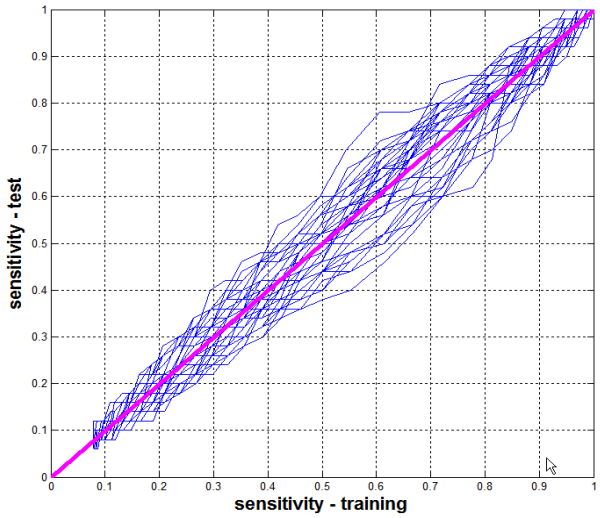


Figure 2.14: Sensitivity of the tested arrays versus the sensitivity of the training arrays. Each line represents one trial of the cross-validation process, while each point on a line represents the sensitivity of the test and the training arrays at the same tuning parameter.

Using MIS, the deviation of sensitivity between the tested and the training arrays is less than 5% in 83% of the time, and less than 10% in 97.5% of the time. For the other algorithms, the deviation is limited to less than 10% for 97%, 96.5%, and 96% of the time using TLRT, UMM, and BPWT, respectively.

The plot also indicates that, the sensitivity is monotonic with the tuning parameter in all cases. That means, if the sensitivity at parameter L_1 is higher than the sensitivity at parameter L_2 in one experiment, then L_1 will always provide a higher sensitivity than L_2 in any other experiment.

The deviation of sensitivity at high and low values is much less than the deviation at the middle of its range. This is an advantage since the experiments are preferred to be run at higher values of sensitivity, which corresponds to a less deviation. Figure 2.15 illustrates the mean and standard deviation of sensitivity under each tuning parameter. It is almost guaranteed that, the deviation of sensitivity between any two experiments is less than 5% if the sensitivity is higher than 80%.

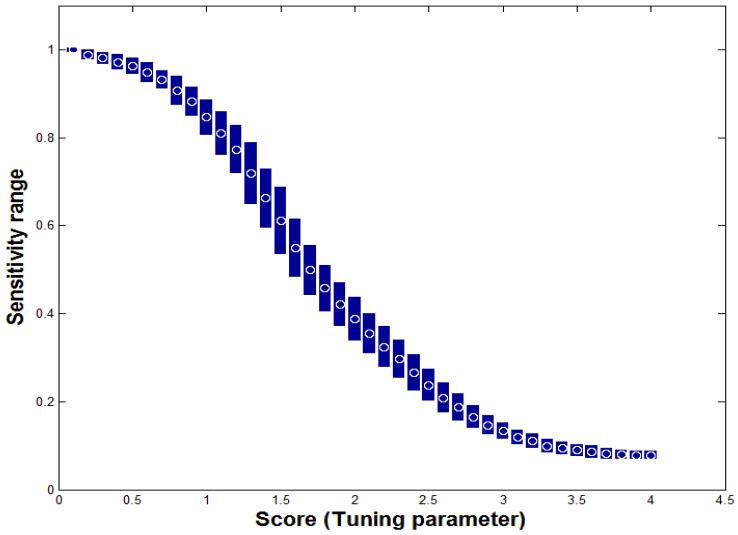


Figure 2.15: Variability of sensitivity under different tuning parameters of MIS. The white circles represent the sensitivity mean μ, and the blue bars cover the distance from μ-σ to μ+σ, where σ is the standard deviation of the sensitivity at each parameter.

The deviation of the false alarm is larger than the deviation of the sensitivity. Figure 2.16 illustrates the results of the cross-validation of the false alarm rates under the same tuning parameters. However, the arrays can be forced to generate very similar false alarm rates if they are scaled to have the same variance. In that case, the generated false alarm rate is similar in all experiments but the sensitivity is significantly different.

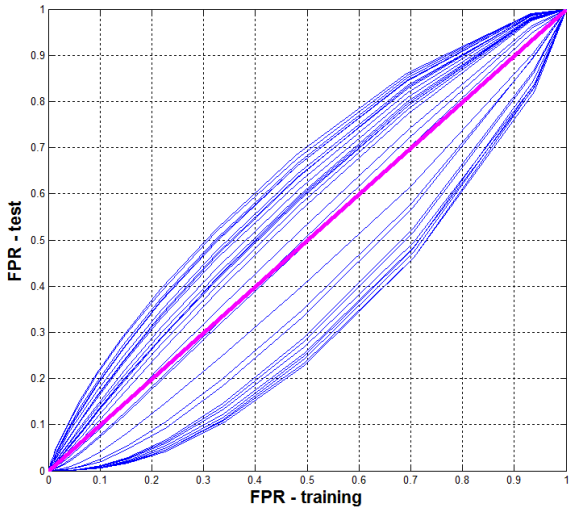


Figure 2.16: False alarm rates of the tested versus the training arrays. Each line represents one trial of the cross-validation process, while each point on a line represents the false alarm rate of the tested and the training arrays at the same tuning parameter.

The conclusion of the experiment is that, the sensitivity of the test-reference arrays is reasonably stable and it has similar values in different experiments. This result is not unexpected since the ratio between different levels of DCN is constant, and the tuning parameter is mostly related to the detected segment's mean [26,36,46]. The $\log_2$ ratio of one gained copy is approximately $\log_2(1.33) \approx 0.42$ in several platforms [36]. We will explain that in details when we discuss the stationarity of the distribution of the microarrays in section 3.8. The variability of the false alarm rate can be justified because it is independent of the DCN variation and of the tuning parameters as well. It is not guaranteed to have the same false alarm rate even if the two arrays belong to the same individuals and were created at the same genotyping lab using the same platform and protocol.

Although, the full ROC is not identical from one experiment to another, a specific range of it can be reproduced with reasonable accuracy. As shown in figure 2.15, the deviation of the sensitivity from one experiment to another decays as the sensitivity reaches to 0 or 1. The same manner happens in the deviation of the false alarm rate as shown in figure 2.17. Therefore, the deviations of sensitivity and false

alarm rate are minimal at certain range of the tuning parameter. As an example, MIS with score = 1 has a false alarm rate bound to less than 7% with sensitivity values bound between 0.8 and 0.92 almost in all cases. The accuracy of reproducing the same ROC points decreases as the value of the score changes from 1. Table 2.3 presents the range of sensitivity and false alarm at selected tuning parameters for the four algorithms.

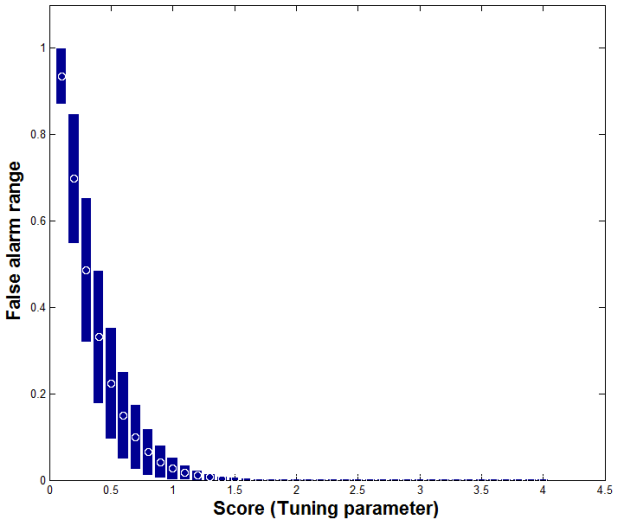


Figure 2.17: The variability of the FPR versus the tuning parameter in MIS. The white circles represent the mean μ and the blue bars cover the distance from μ-σ to μ+σ, where σ is the standard deviation of the false alarm values at each tuning parameter.

Algorithm	Selected tuning parameter	Range of TPR	Range of FPR
MIS	score = 1	0.80-0.92	0-0.07
UMM	least state = ±0.6	0.62-0.72	0-0.05
TLRT	mean = ±1.9	0.76-0.92	0-0.08
BPWT	universal threshold = 2.5	0.82-0.94	0-0.09

Table 2.3: Range of sensitivity and false alarm rate at selected tuning parameters.

2.8 Conclusions

The main conclusion of the comprehensive comparison is that the problem of analyzing CGH arrays is fully suited for linear algorithms. Our simple algorithms MIS and TLRT outperformed all other algorithms in the comparison. Most of the existing algorithms are much more sophisticated in theory and much slower in processing than TLRT and MIS.

The underlying solution of the DNA microarrays is very smooth with a very few transitions compared to the data size. And about one half of the solution's segments are known to be on the baseline. The quadratic algorithms, which represent the vast majority of algorithms in the literature, are redundant to estimate smooth functions as it is shown in table 2.2. HMM is widely considered in pattern recognition problems where the model switches more frequently among the hidden states as opposed to the data of CGH microarrays where the transitions are extremely sparse. Also the clustering algorithms are pivotal in machine learning and image processing applications. But employing such complicated and time-consuming algorithms in analyzing the DNA microarrays is overemphasizing the problem and does not necessarily provide good solutions.

The ROC curves are different from one experiment to another even when the same test is applied with the same parameters at another data from the same individuals. Luckily, the deviation of ROC is concentrated at non-interesting regions where the sensitivity is low or the false alarm rate is high. The deviation of the ROC curves at high sensitivity and low false alarm rate is fairly low which guarantees that the ROC curves can be reproduced with a reasonable accuracy.

Chapter 3

A Statistical Model For Genomic Hybridization Experiments

3.1 Introduction

The genomic hybridization experiment is a very popular tool to read the DNA copy number or the gene expression in normal or abnormal cells. Thousands of experiments are conducted every year to accumulate the knowledge about the human genome. Over the years, many algorithms were developed to identify the variation of the DNA copy number using the conventional two-channel approach. A crucial aspect of the analysis is to develop a deterministic model that fits the distribution of the data accurately. The developed model is the foundation that drives the process of all methods. The power of detection and the limitations of any method are all dependent on the assumed model.

Although the model identification is very critical to the performance, it has not gained the attention it needs. Commonly, the two-channel methods consider the independent and identically distributed Gaussian model and start the analysis based on that assumption [5-25]. Several other models were proposed [46-72] but their main focus was on modeling the comparative profile ($\log_2$ ratios) instead of modeling the raw

intensities of each channel per se. They present their single-channel models just to explain how to address the analysis of the $\log_2$ ratios. The problem of finding an accurate model that fits the distribution of single-channel microarrays has yet to be satisfactorily explored.

In a real genomic hybridization experiment, several sources contribute to the final measurement of the intensity of each probe. Some of these sources are biological and some of them are systematic. The biological sources include, but not limited to, perfect hybridization, cross hybridization, missing targets, and GC contents. The systematic sources include, but also not limited to, background effect, scanner's bias, and fragment length through the performance of the PCR. A robust model should take most or all these factors into consideration.

The main contributions of this chapter are 1) a novel *Quantile-based Perfectly Isolated* model (QPI) which isolates the desired distribution from a mixture of non-homogenous distributions using the observations' quartiles, 2) a *Universal Threshold Adjustment* model (UTA) to remove the bias of the imperfect scanner, 3) GCNORM, a normalization model for the GC content, 4) FLNORM, a novel source-based normalization model for the fragment length's bias, and 5) a proof of the stationarity of the microarrays and its impact on the computations.

In section 3.2, we give an introduction to the Genome-Wide Human SNP Array 6.0 produced by Affymetrix and state the problem. In section 3.3, we discuss several models of the distribution and present the QPI model. In section 3.4-3.6, we present UTA, GCNORM, and FLNORM to remove the bias of scanner, GC content, and fragment length. In section 3.7, we verify the QPI model and show results using real-world data from the international Hapmap project. In section 3.8, we demonstrate that the genomic hybridization process is stationary and we emphasize the impact of that result on the model's accuracy and computational burden.

3.2 Genome-Wide Human SNP Array 6.0 - Affymetrix

Here we give a detailed description of the design and layout of genome-wide human SNP 6.0 arrays manufactured by Affymetrix. The experiment is conducted on a chip which is ½ inch × ½ inches (1.28cm x 1.28cm) comprising 6,892,960 probes sorted in 2572 rows and 2680 columns. Each probe consists of millions of identical short sequences complementary to their targets' sequences. The 6,892,960 probes target a total of 1,856,069 sites on the human genome. The total number of probes is higher than the total number of designated targets on the genome. The targets of each genomic site are captured by one probe, six probes, or eight probes.

After scanning the chip, an image is created to provide the raw intensity of each probe. Each probe of the chip is represented by one pixel in the image. The intensity measurements and the (x,y) coordinates of each pixel are embedded in a CEL file. The scheme that maps the image coordinates into their 1,856,069 genomic sites is embedded in an SPF file. The only explanation provided by Affymetrix for the reason of not combining all the information in one single file is the size limitation.

The 1,856,069 genomic sites of the genomic profile are divided into three main groups. The first group consists of 906,600 single nucleotide polymorphism (SNPs) sites: 869,481 of them are located at chromosomes 1 to 22 (autosomes), 36,862 sites are located at chromosome X, and 257 sites are located at chromosome Y. 796,045 of the SNP sites are represented by 6 probes on the chip (3 for each allele) and 110,555 SNP sites are represented by 8 probes (4 for each allele).

In the literature and the software packages, the intensity of each SNP site is calculated as the average of the intensities of the corresponding 6 or 8 replicate probes. Several linear and nonlinear averaging tools are used to calculate the needed intensity. Averaging the replicate probes is not accurate because it makes the average intensity of a SNP probe equal to one sixth or one eighth of the average intensity of a CN probe. And since the CN and SNP sites are contiguous along the genome, the distribution of any segment of the genomic profile will be a mixture of three different populations.

Furthermore, the distribution of the SNP probes is not homogenous because of the homozygosity and the heterozygosity of the alleles. The homozygous targets bind to one allele probe-set while the heterozygous targets bind to the two replicate sets equally. Figure 3.1 illustrates the distribution of all SNP probes and specifies its components.

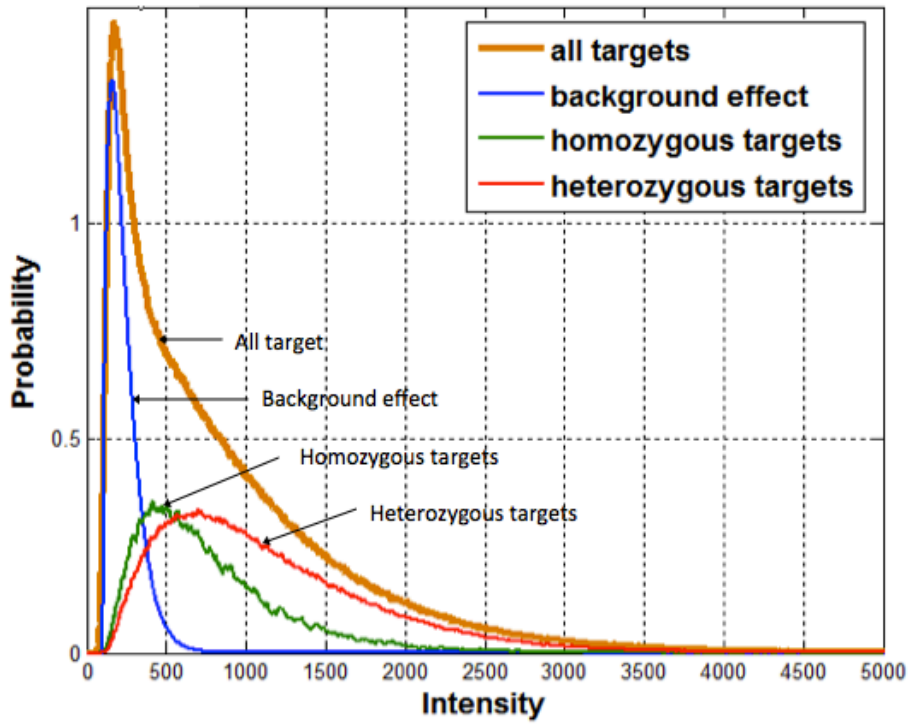


Figure 3.1: The intensity distribution of SNP probes

Therefore, the sum of the intensity of the replicate probes of each SNP site is supposed to be used instead of any other averaged measurement. Using non-linear averaging measurements of the SNP probes is a very common mistake in the software packages.

The second group of the genomic profile consists of 945,826 non-polymorphic sites called “copy number” CN sites. 888,043 of them are located at chromosomes 1 to 22, 49,201 sites are located at chromosome X, and 8,582 are located at chromosome Y. Each CN site is represented by a single probe on the chip.

The third group of the genomic profile consists of 3,469 sites that are used to verify the sample's identity. The 3,469 markers are represented non-uniformly by 81,744 probes on the chip. The rest of the probes on the chip do not have targets in the assay. These 204,680 are meant to capture the background effect to be used in the analysis. All statistics are presented in table 3.1.

In general, the probes are distributed on the chip according to their target's types. The CN probes form a plus sign (+) in the middle of the chip and divide the rest of it into four rectangular shapes as illustrated in figure 3.2. The CN probes occupy the columns from 1245 to 1436 and the rows from 1193 to 1380. The SNP probes and the majority of the non-targeting probes are distributed almost uniformly in the upper and the lower right and left sides of the image. The intensities of CN probes are, in general, greater than the intensities of SNP probes since all CN targets bind to single probes while the targets of SNP probes are distributed among 6 or 8 replicates.

Each probe on the chip has an identifier (probe ID) and a serial number. The serial numbers count from 1 to 6,892,960 starting from the upper left corner and increasing as they go to the right side. When the counter reaches the end of a row, it continues from the left side of the next row. The probe-IDs are chosen arbitrarily without following a specific sequence. The replicate probes that capture targets from the same SNP genomic

Chr.	**CN**		**SNPs**				**Non-targeting**	
			Rep. by 6-probes		**Rep. by 8-probes**			
	Array	**Chip**	**Array**	**Chip**	**Array**	**Chip**	**Array**	**Chip**
1-22	888,043	888,043	766,210	4,597,260	103,271	826,168	3,469	286,424
X	49,201	49,201	29,578	177,468	7,284	58,272		
Y	8,582	8,582	257	1,542	0	0		
total	945,826	945,826	796,045	4,776,270	110,555	884,440	3,469	286,424

Table 3.1: Statistics of CN and SNP probes on the GWS6.

site share the same probe ID but not the same serial numbers. The serial number is used to localize the probe on the chip while the probe ID is used to map the image intensities into the genomic sites. The length of all probes is 33-mer with 16 bps on each side of the center. The center of the SNP probes is different according to the allele. The length of targeted fragments ranges from 0 to 50,000bp but it concentrates mainly in the range from 200 to 2000bps. The GC content of each probe is measured as the percentage of G and C bases in a 500,001bp window centered at the genomic site targeted by that probe.

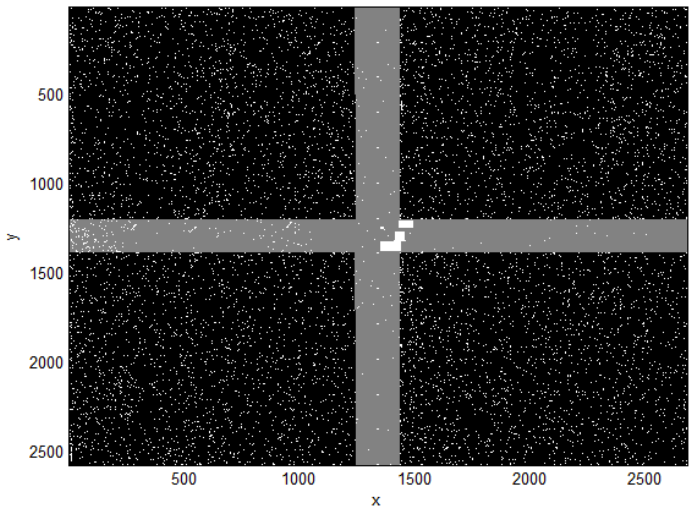


Figure 3.2: The probes' layout on a typical GWS6 chips. Black pixels are SNP probes, gray pixels are CN probes, and white pixels are non-targeting probes.

For clarity, we will be referring to the group of CN probes by the notation C and the group of SNPs by S. The notations are followed by either A, X, or Y to refer to the chromosome's type: autosomes, X or Y, respectively. For example, CX is the group of CN probes whose targets are located in the X-chromosome, whereas SA is the group of SNP probes whose targets are located in the 22 autosomes. We will add a sub-notation 6 or 8 to separate the SNP probes based on the number of the replicate probes allocated for every site. We also will refer to the intensities in the image as "I" while the intensities on the genomic profile sites are referred to as Y. $I_{(x,y)}$ is the intensity at the point (x,y) in the image and Y_i is the intensity at the site i on the genome.

3.3 Quantile-based Perfectly-Isolated model (QPI)

One of the central steps in the detection process of any particular method is to infer the probability distribution function of the observations if it is not given. The detection criteria using the likelihood test under the Bayesian formulation or Neyman-Pearson theory are fully dependent on the inferred distribution. Accurate estimation of the distribution is very essential for accurate detection performance. Furthermore, accurate knowledge of the distribution gives insight into the limitation of the detection performance and whether the solution is realistic or not. In this section, we present the QPI model which precisely fits the distribution of the DNA microarrays. The model will be the cornerstone of the analysis in chapter 4.

3.3.1 related work

Several models were proposed in the literature to fit the distribution of the microarrays intensities [26,36,46-72]. These several attempts emphasize the importance of using an accurate model of the distribution since it is the main basis of the analysis. We will briefly discuss the existing models and show the deviation between them and the actual distribution. We will highlight a few points before starting the discussion.

First, there is a consensus in all references that the probe's intensity level is proportional to the amount of DNA targets that bind to it. The total number of targets depends on the total number of DNA molecules and on the copy number level. The effect of the copy number variation on the intensity can easily be demonstrated by comparing the intensity mean of a male X-chromosome with the intensity mean of other autosome. In several references, [26,36,46-49], the relationship is specifically linear. In [48], the authors conducted a wide experiment to compare the intensity mean of several known levels of the X-chromosome where the copy number ranges from 0 to 1000. The authors concluded that the relationship between the intensities and the copy number fits a straight line not passing through the origin. The effect of the number of molecules is proven in [50] and the intensity depends on it linearly as well.

Second, all microarrays have similar distributions. The distribution is un-modal with a short tail on the left side and a very long tail on the right side as shown in figure 3.3.

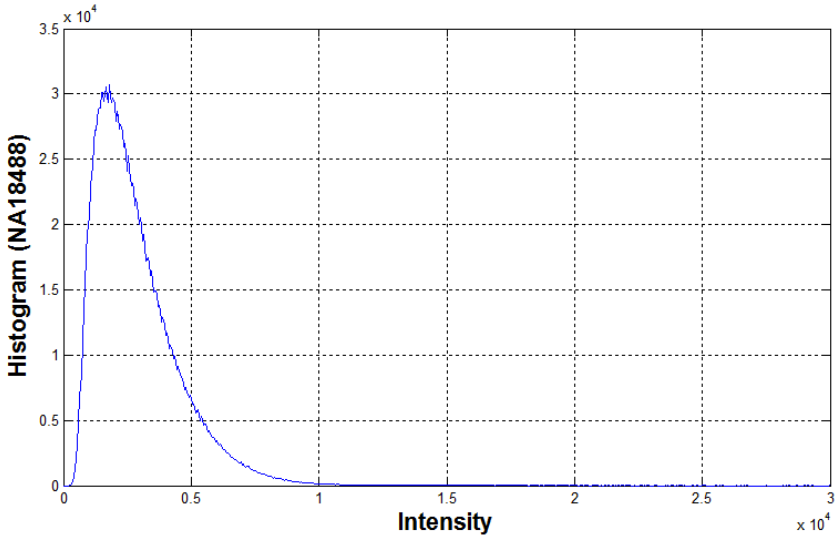


Figure 3.3: CN intensities histogram of sample NA18488

Third, the binding process is affinity-based which means that a target only binds to a probe if they have a partial or a full complementary sequence. Each probe attracts specific targets from the assay where some of them match its sequence perfectly and some of them match its sequence partially. The latter is widely known as the "cross hybridization". The affinity base implies that, the cross hybridization component is dependent on the sequence [51],[52]. In other words, the cross hybridization component is specific rather than stochastic. This component emerges from poorly designed probes because of the low specificity that the probes might have. The conclusion is that, the cross hybridization (the noise) is highly correlated to the probes' design. That can be demonstrated by measuring the cross correlation among independent samples. Figure 3.4 illustrates the cross-correlation of group "EPODE" from the International Hapmap Project. The values of the cross correlation range from 0.75 to 0.95 and that indicates a significant correlation among samples.

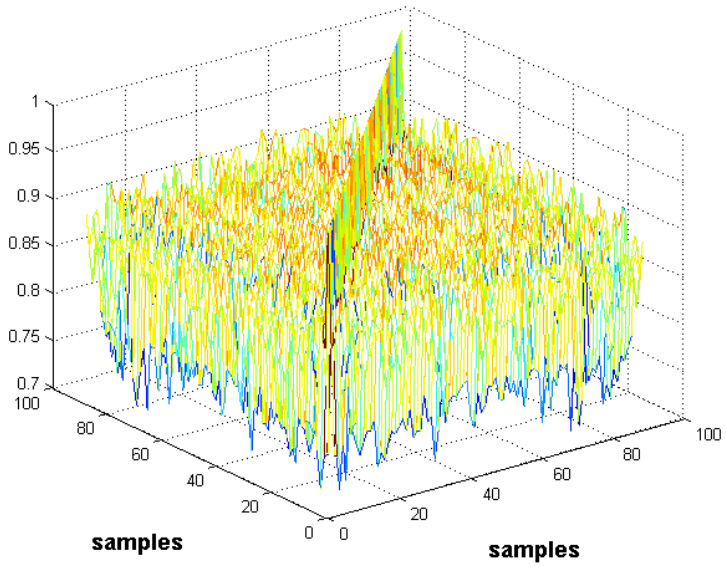


Figure 3.4: Cross correlation among "EPODE" sample of the IHP

Fourth, there is a debate in the literature about which component is stronger: the specific hybridization or the cross hybridization components. The authors in [53] assume that the power of the specific hybridization component is much stronger than the cross hybridization component whereas the authors in [54] contradict that conclusion by adopting the opposite result. We will show that the specific hybridization in the autosomes is twice stronger than the cross hybridization component.

Finally, there are some factors that occur during the kinetic of the experiment and they only effect the duration of the hybridization process. Some of these factors are: binding rate, target concentration [58], detachment rate, gas constant, and the temperature [59]. We will not explore any of these factors since the targets of any experiment are usually given enough time to hybridize and that allows each probe to reach to its steady state intensity.

Most of the proposed models in the literature consider additive noise components and some consider multiplicative components. The multiplicative components convert to additive components in the $\log_2$ space.

The early models [46,47] were limited to specifying the relationship between the intensity mean and the number of DNA copies in a specific genomic site. All models agree that the relationship fits a straight line not passing through the origin. The model in [47] suggests that the slope of the straight line is controlled by the specific hybridization while the intercept is controlled by the cross hybridization component. The model in [46] specifies that the intensity mean corresponding to n DNA copies is equal to:

$$I_n = 0.27 + 0.37.n$$

The slope in the ideal case is 0.5 since the human genome has two copies at each site. But the slope deviates from that value because of the cross hybridization which reduces the slope to less than 0.5. Slopes that are close to 0.5 correspond to less error. The model does not provide any more details about the distribution's statistics other than the mean. The ratio between the specific and cross hybridization components in this model is (0.27/0.37)≈0.73.

A similar model is presented in [26] using a similar experiment. Their result defines the intensity mean as:

$$I_n = 0.44 + 0.28n$$

The slope is also not equal to 0.5 and they used the same explanation in [46] to justify the deviation.

Following the same approach, a wide comparison is conducted in [36] to measure the slope in different platforms. In all cases, the slope is bound between 0.22 and 0.42 with a mean of 0.38.

The author of [60] proposes a multiplicative model of the intensity level as:

$$I[i] = \alpha(\gamma \times A[i] \times C[i] + \beta)$$

Where α is a multiplicative system noise, γ is labeling noise with mean = 1. A is a performing factor for the probe, C is the copy number divided by 2, and β is an additive background noise. The model in [61] suggests that the intensity is measured as:

$$I = \mu \times A \times T \times e$$

Where μ is the expected value of the intensity, A and T are the effect of the chip's design and the effect of the sample, respectively. The term e is a correlated noise. Similar approaches are presented in [62], [63], and [64].

The previous models reveal only the relationship between the intensity mean and the copy number, but they don't provide more details about other statistics of the intensity distribution or how to estimate them. The following models shed more light on these regards.

The model PDDN in [65] and [66] consists of two components and it determines the intensity as:

$$I = \frac{N}{1 + \exp(E)} + \frac{N^*}{1 + \exp(E^*)} + B$$

Where N is the population of the perfect matching targets, and N^* is the population of the cross hybridization targets. E and E^* are factors representing the free energy for formation of the specific and cross hybridization targets, respectively. B is a uniform background. The model is interesting since the sum of the two components is uni-modal even though each one of them is a monotonically decreasing function. Intuitively, this assumption can not be true since the distribution of the perfect hybridization is not monotonic. Rather, it includes a local maximum value closer to the left side of its domain. Furthermore, we will prove that the distribution of the cross hybridization component is not monotonic either.

The model in [49] assumed that the distribution of a single specific target is binomial. And since the number of targets is usually large, the distribution of the specific hybridization tends to be Gaussian. Another Gaussian component is assumed to represent the intensity of the cross hybridization. This model is not realistic since the distribution, as illustrated in figure 3.3, is far from being symmetric.

A very interesting model in [67] suggests the following distribution to fit the distribution

$$I = \alpha + \mu.\exp(\eta) + \varepsilon$$

Which implies that $\log[I - \alpha] \approx \log(\mu) + \eta$

μ is the mean intensity value, α is a small constant to represent the mean background intensity, ε and η are Gaussians random variables with ε << η. Therefore, the distribution of this model is almost Gaussian in the log space and that is the main concept of the widely accepted Gaussian model for the $\log_2$ ratios. However, the Gaussian model is not precise to fit the distribution of a single channel. The distribution of a single channel decays slower than the Gaussian distribution on their left sides as shown in figure 3.5. A similar model is suggested in [68] also.

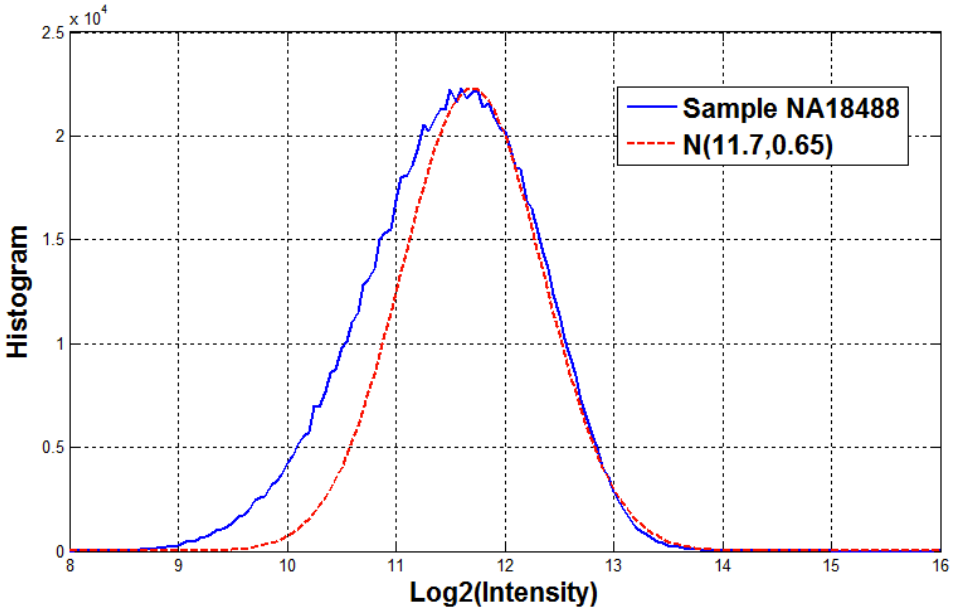


Figure 3.5: Histogram of log2 ratios and normal distribution

In [53], the distribution of the raw intensities is modeled as the sum of two components: significant Gaussian component to represent the specific hybridization and a slight exponential component to represent the cross hybridization. The two components can not fit the distribution of the microarrays unless the cross hybridization component is at least ten times stronger than the perfect hybridization. And that requirement never exists in any known platform and it also contradicts the first assumption of the model which assumes the cross hybridization component to be relatively smaller than the perfect hybridization.

3.3.2 The QPI model

There is a common fault in most existing models where the noise is assumed to be homogenous. The models assume that the observations are drawn from the same distribution while, in fact, they are not.

The main biological source of intensity is the *actual* amount of copy number in the interrogate sites. The intensity distribution of probes that have n DNA copies in the assay is different than the distribution of probes that have m DNA copies where $m \neq n$. That means the observations are drawn from *n* different distributions, each one of them represents a different level of copy number with one highly dominating component representing the diploid event (two copies in an ideal DNA sequence).

Another aspect in the analysis is the heavy tail on the right side of the distribution which generates significant amounts of outliers. Some outliers reach to the saturation level (65,536) and they are responsible of generating many false calls. More than 3% of the observations fall beyond a span of 3 standard deviations from the mode compared to 0.27% in the Gaussian distribution. That means the raw observations of the mixture are not compatible to be used directly to estimate the mean and the standard deviation of the diploid distribution because of the existing duplications and deletions. Therefore, we need to create a model that is insusceptible to the outliers and to the variation of the copy number.

By inspecting the observations, we see that all intensities are strictly greater than zero and they have an upper bound, 65,536. The upper bound guarantees that the observations' mean is bound. The strict lower bound guarantees that the mean is strictly positive. And the two bounds guarantee that the log values and their mean are bound as well. We also find out that the distribution is uni-modal and positively skewed. All these characteristics exist in the gamma distribution.

The selection of the gamma distribution to model the observation is a very reasonable choice for two reasons. First, the gamma distribution exhibits the highest entropy of any non-uniform distribution of strictly positive values [41]. It is preferred to use the distribution that conveys the least amount of certainty about the data. Second, the

gamma distribution involves several famous distributions as special cases such as the Gaussian, exponential, and Erlang distributions. Furthermore, the gamma distribution is similar to some heavy tailed distributions such as beta and lognormal distributions. Therefore, we will start with the assumption of having a gamma distribution and will check if any special cases exist or not.

The gamma distribution is defined by two parameters: a shape parameter, k, and a scale parameter, θ. And the probability density function is:

$$f(x,k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

Where Γ(k) is gamma function and it is defined as:

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$$

Several approaches have been presented in the literature to estimate the Gamma distribution's parameters using the maximum likelihood [44]. However, the maximum likelihood estimators are not accurate to estimate the parameters in the microarrays since the distribution is not homogenous. We suggest using the quantiles because they are statistically robust against the outliers and are less affected by the copy number variation if one of the mixture's components is highly dominating the others. Here we suggest using the quartiles Q_1, Q_2, and Q_3, of the Gamma probability function f(x,k,θ) which are defined as:

$$\begin{aligned} \int_0^{Q_1} f(x,k,\theta)dx &= 0.25 \\ \int_0^{Q_2} f(x,k,\theta)dx &= 0.5 \\ \int_0^{Q_3} f(x,k,\theta)dx &= 0.75 \end{aligned} \qquad (3.1)$$

Hence, Q_2 is equivalent to the median. We will present approximated closed form equations of the quartiles based on the parameters. The quartiles can be inferred from the observations and then, the parameters can be inversely estimated using the closed form equations.

Before deriving the closed forms, a constant bias existing in all arrays needs to be considered as well. The bias Δ causes a linear shift to the intensities and it must be estimated to correctly extract the parameters.

QPI model:

The scaling property of the gamma distribution states that: $\Gamma(k,\theta) = \theta.\Gamma(k,1)$. That means the quartiles of any gamma distribution are equal to the quartiles of $\Gamma(k,1)$ multiplied by θ. i.e. $Q_i(k,\theta) = \theta.Q_i(k,1)$. Therefore, it is sufficient to derive an approximation of the closed forms of $\Gamma(k,1)$ and it can be scaled to fit any other distribution. We suggest the following approximated closed forms of the quartiles.

$$\begin{aligned} \hat{Q}_1(k,1) &= 0.2875 + 0.6746(k-1)^{1.095} \\ \hat{Q}_2(k,1) &= 0.6930 + 0.9853(k-1)^{1.007} \\ \hat{Q}_3(k,1) &= 1.3861 + 1.3056(k-1)^{0.954} \end{aligned} \qquad (3.2)$$

Considering the scaling and the shift, the final approximation is:

$$\left\{ \begin{aligned} \hat{Q}_1(k,\theta) &= \Delta + [\,0.2875 + 0.6746(k-1)^{1.095}\,].\theta \\ \hat{Q}_2(k,\theta) &= \Delta + [\,0.6930 + 0.9853(k-1)^{1.007}\,].\theta \\ \hat{Q}_3(k,\theta) &= \Delta + [\,1.3861 + 1.3056(k-1)^{0.954}\,].\theta \end{aligned} \right\} \qquad (3.3)$$

The relative error of this estimator is shown in figure 3.6. The relative error for $\hat{Q}_i$ is defined as:

$$Error\% = 100 * \frac{|Q - \hat{Q}_i|}{Q}$$

The plot also presents the relative error of Banneheka closed form of the median [69] which is equivalent to Q_2 in (3.2). The relative error is less than 0.5% for $1.5 \le k \le 6.5$. We will show that, the shape parameter in the microarrays distribution always falls within this range.

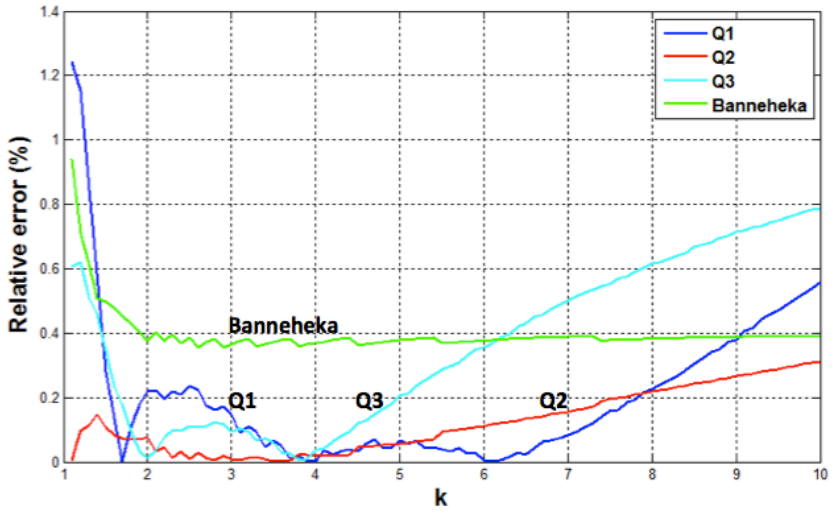


Figure 3.6: Relative error of the closed forms of the quantiles. The dashed line is the relative error of the closed form of [69].

The group of equations in (3.3) can used inversely to estimate the shape and scale parameters and the constant shift. Q_1, Q_2, and Q_3 are obtained directly from the observations as defined in (3.1). The solution of the shape parameter, k, can be obtained by solving the following polynomial whose exponents are not integers. The solution of the polynomial can be obtained numerically using any choice of the household's methods.

$$0.6746(Q_3 - Q_2)(\hat{k} - 1)^{1.095} - 0.9853(Q_3 - Q_1)(\hat{k} - 1)^{1.007} + 1.3056(Q_2 - Q_1)(\hat{k} - 1)^{0.954} = 0.4055Q_3 - 1.0986Q_2 + 0.6931Q_1 \quad (3.4)$$

And the scaling factor can be estimated as:

$$\hat{\theta} = \frac{Q_2 - Q_1}{0.4055 + 0.9853(\hat{k}-1)^{1.007} - 0.6746(\hat{k}-1)^{1.095}} \quad (3.5)$$

And finally, the constant shift is estimated as:

$$\hat{\Delta} = \hat{Q}_1(\hat{k}, \hat{\theta}) - \left[0.2875 + 0.6746(\hat{k} - 1)^{1.095}\right].\hat{\theta} \quad (3.6)$$

3.3.3 Modeling the intensity distribution of gains and losses:

As we mentioned earlier, the distribution of the diploid fragments consists of two major components and two minor components. The major components are the true specific hybridization and the false cross hybridization. The minor components are the background effect, which is random in nature, and a constant bias. The background effect is relatively smaller than the cross hybridization component. The constant bias Δ is just a linear shift to the intensities and we explained how to estimate its value in the QPI model. Intuitively, the value of Δ is the same for all distributions within the same array ($\Delta_{diploid} = \Delta_{gain} = \Delta_{loss} = \Delta_{CA} = \Delta_{SA} = \Delta_{CX} = \Delta_{SX} = \Delta$). In the rest of this section, we will assume that the constant bias Δ has been subtracted from the raw intensities. We also will define F as the sum of the cross hybridization and the background effect (FH = CH + BC) and define TH as the true specific hybridization. TH and FH are independent random variables with non-identical distributions. We will use the subscript i to refer to the copy number that the intensities correspond to. The ideal case is when i = 2.

1 copy loss (single copy detected):	$H_1 = TH_1 + FH \sim \Gamma(k_1, \theta_1)$
No variation (two copies detected):	$H_2 = TH_2 + FH \sim \Gamma(k_2, \theta_2)$
1 copy gain (three copies detected):	$H_3 = TH_3 + FH \sim \Gamma(k_3, \theta_3)$

The distribution of FH is supposedly identical in all events (H_i, i = 0,1,2,…), since it only depends on the chip's design. The events are only different in their specific hybridization components TH_i which depends on the actual amount of DNA copy number. TH_0 is equal to zero since there are no targets available in the assay because the two copies are already lost. The total number of targets from a single genomic site with i copies is equal to (i/2) multiplied by the total number of targets from the same site if it was diploid. Therefore, all TH_i's components belongs to the same family of TH_1 with scaling factors i. Using the scaling property of the gamma distribution: if $TH_1 \sim \Gamma(k,\theta)$ then $TH_i \sim \Gamma(k, i\theta)$. Clearly from the equations, the mean of H_1 is not equal to 0.5 of the mean of H_2 due to the existence of FH. This attribution was mentioned non-definitively in [46] and we assure its validity.

$$
\begin{aligned}
&\text{1 copy loss (single copy):} && H_1 = TH_1 + FH \;\sim\; \Gamma(k_1, \theta_1) \\
&\text{No variation (two copies):} && H_2 = 2TH_1 + FH \;\sim\; \Gamma(k_2, \theta_2) \\
&\text{1 copy gain (three copies):} && H_3 = 3TH_1 + FH \;\sim\; \Gamma(k_3, \theta_3)
\end{aligned} \tag{3.7}
$$

And so on. The parameters k_2 and θ_2 can be estimated using the data of the autosomes as explained in the QPI model. And k_1 and θ_1 can be estimated from the X and Y-chromosomes for male samples. We will continue this section considering only male samples and will generalize the model in section 3.8 to include female samples.

From the equations 3.7, we can infer the following statistics:

$$
\begin{aligned}
E[TH_1] &= E[H_2] - E[H_1] = k_2\theta_2 - k_1\theta_1 \\
E[FH] &= E[H_1] - E[TH_1] = 2k_1\theta_1 - k_2\theta_2 \\
\Rightarrow E[H_i] &= E[TH_i] + E[FH] \\
&= (i-1)k_2\theta_2 + (2-i)k_1\theta_1
\end{aligned}
$$

Also:

$$
\begin{aligned}
Var[H_1] &= Var[TH_1] + Var[FH] = k_1\theta_1^2 \\
Var[H_2] &= 4.Var[TH_1] + Var[FH] = k_2\theta_2^2
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
Var[TH_1] &= (k_2\theta_2^2 - k_1\theta_1^2)/3 \\
Var[TH_i] &= i^2(k_2\theta_2^2 - k_1\theta_1^2)/3 \\
Var[FH] &= (4k_1\theta_1^2 - k_2\theta_2^2)/3 \\
\Rightarrow Var[H_i] &= Var[TH_i] + Var[FH] \\
&= [(i^2-1)k_2\theta_2^2 + (4-i^2)k_1\theta_1^2]/3
\end{aligned}
$$

And finally, the shape and scale parameters for any event H_i can be estimated as:

$$\left.\begin{aligned} k_i &= 3\,\frac{[(i-1)k_2\theta_2 + (2-i)k_1\theta_1]^2}{(i^2-1)k_2\theta_2^2 + (4-i^2)k_1\theta_1^2} \\ \theta_i &= \frac{(i^2-1)k_2\theta_2^2 + (4-i^2)k_1\theta_1^2}{3[(i-1)k_2\theta_2 + (2-i)k_1\theta_1]} \end{aligned}\right\} \qquad (3.8)$$

By finding the values of k_i's and θ_i's, the distributions of all events are known and the model is complete. The remaining is to employ this information into the detection technique of chapter 4 to infer the real statuses.

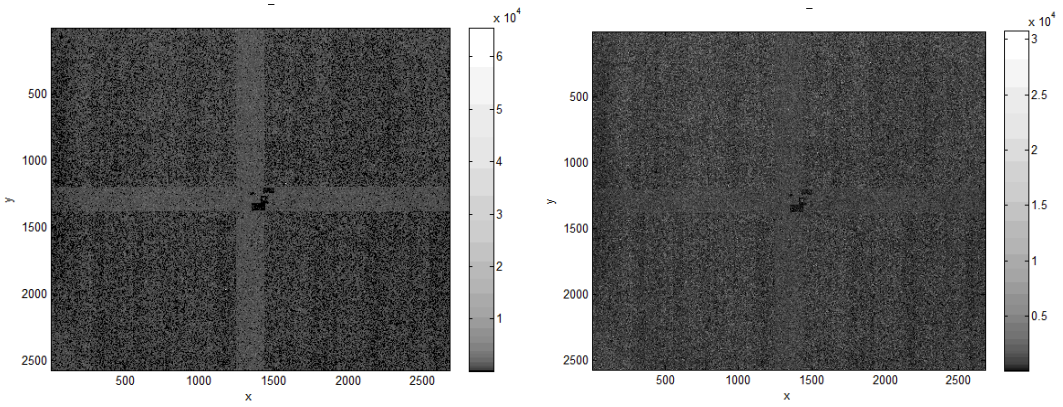
3.4 Removal of Systematic Bias

In this and the following two sections, we will discuss three sources of bias in the DNA microarrays: imperfect scanner, GC contents, and fragment lengths. These sources are totally independent of the biological status (gain, loss, or normal) of the DNA copy number. Each source of bias requires a different normalization process to be removed. The existing normalization models are mainly developed for the two-channel approaches. These models either normalize the arrays to each other or normalize their $\log_2$ ratios with respect to the source of bias. The models we present here follow the other direction where each array is normalized within itself. We chose the single-channel approach because the normalization process can be applied directly into the observations. In the two-channel approaches, the normalization process is applied onto normalized $\log_2$ ratios instead of the original values.

As we mentioned before, the mean intensity of a probe is directly proportional to the number of its targets in the assay and inversely proportional to the number of the identical replicate probes. Therefor, the mean intensities of CA, CX, CY, SA_6, SA_8, SX, and SY are different from each other. And other statistical measurements such as

median, mode, and standard deviation are directly proportionate to the mean of each population. Therefore, the first step in the process is to scale all these groups to have the same arithmetic mean. Figure 3.7 depicts a typical microarray image before and after the mean-normalization step. We chose the mean of CA to be the reference:

$$\left\{\begin{array}{rcl} SA_6 & \rightarrow & \mathrm{SA}_6 \times \mathrm{E[CA]/E[SA_6]} \\ SA_8 & \rightarrow & \mathrm{SA}_8 \times \mathrm{E[CA]/E[SA_8]} \\ SX & \rightarrow & \mathrm{SX} \times \mathrm{E[CA]/E[SX]} \\ SY & \rightarrow & \mathrm{SY} \times \mathrm{E[CA]/E[SY]} \\ \mathrm{CX} & \rightarrow & \mathrm{CX} \times \mathrm{E[CA]/E[CX]} \\ \mathrm{CY} & \rightarrow & \mathrm{CY} \times \mathrm{E[CA]/E[CY]} \end{array}\right\} \quad (3.9)$$



(a) (b)

Figure 3.7: Typical images of Affymetrix 6.0 arrays. (a) before and (b) after mean normalization.

3.4.1 Introduction to Scanner bias

Figure 3.7 depicts the output image of a male individual's DNA sample. As mentioned earlier, the pixels that form a large "+" sign in the middle of the image represent the CN probes which are supposed to have higher intensities in any sample. The contrast between the CN and the SNP probes becomes vague after the mean normalization step.

In both images, it is easy to notice bright and dark areas across the image and these areas extend vertically. The image appears like it consists of non-homogenous vertical stripes. There is no biological meaning behind this phenomenon and it exists with high amounts of correlation. Most of the bright (and the dark) spots remain bright (or dark) in all images. The process of normalizing one image to another does not remove this noticeable bias since the bias exists in both images. That is one of multiple pitfalls of the two-channel approach which we try to avoid in this work. The source of the contrast is certainly the scanner since each stripe contains uniform mixtures regarding all physical and biological features such as chromosome's type, position, fragment length, GC content, enzyme's type, etc. The only difference among stripe is their spatial locations on the chip.

3.4.2 Related work

The normalizing methods, whether they are applied at the 1-D genomic profile or at the 2-D image, consist of two main steps: detecting the bias, and removing it. The first step can be performed using a 2-D moving average window. The window needs to be large enough to reduce the noise and small enough to preserve the change and not over-smoothen it. Moving mean and moving median windows are frequently used [70] in the literature. The moving median window is more robust to the outliers but it terminates at the boarders between the CN and the SNP probes because the CN and SNP observations belong to different distributions even if they share the same mean intensity.

Another method, which is widely used as well, is Loess regression. It is not robust against the outliers and it is very strenuous in terms of computations. Loess requires 5.8 hours to analyze one image of GWS6 compared to 4 minutes required by the moving median (about 90 folds), and 2.5 minutes by the moving average window (about 135 folds). The computation times were measured using Matlab on a 2.66GHz machine with 48G random access memory (RAM). It is clear that Loess is not properly suited for huge data sets. The Turkey's Weight averaging window is also very expensive in computations and inefficient to be used at large data sets.

The averaged image carries information about the *local* bias at each pixel [70]. Some methods classify the local biases using unsupervised clustering algorithms where each cluster indicates *regional* bias for the included pixels [71]. In both ways, the bias is reduced by dividing the pixels' intensities on their local or regional bias. For more details, we refer the reader to [70].

3.4.3 Universal-Threshold Adjustment (UTA) Algorithm

By looking at figure 3.7, it is clear to notice the discontinuity of the bias between the bright and the dark areas. The contrast of the intensities occurs abruptly, not gradually, and that supports the regional bias approaches over the local bias ones. The main question in the analysis is: into how many clusters should the local bias values be classified? The unsupervised clustering methods seem to over-segregate the results because of the heavy tailed distribution. We found out that considering two distinct clusters is adequate to eliminate the scanner's bias and remove the contrast among the intensities. Therefore, the clustering process must be supervised by forcing the data to cluster into only two clusters: dark and bright. The model consists of three steps.

[step 1]: Outliers elimination

Since the arithmetic mean is not robust to outliers, an additional step to eliminate the effect of the outliers is certainly needed. Here we use a coarse threshold defined as τ multiplied by the standard deviation of the observations where τ is a real number. Any intensity $I_{(x,y)} > \tau$ is replaced by the image's mean.

[step 2]: Image smoothing and edge detector

A square moving average window is applied at the raw intensities $I_{(x,y)}$ to generate a smoothened image S. $S_{(x,y)}$ is equal to the arithmetic mean of the intensities included in a square window centered at (x,y). The window is trimmed near the edges because the intensities are not defined beyond the image's boundaries.

The distributions of the dark and bright clusters are unknown and hence, there is no applicable likelihood function to separate the observations. Therefore, the smoothened pixels are declared Dark or Bright according to the rational test:

$$S_{(x,y)} \underset{Bright}{\overset{Dark}{\lessgtr}} \eta \qquad (3.10)$$

Where η is a universal threshold for the image.

[step 3]: Calculating the universal threshold

The universal threshold can be estimated as the best threshold that segregates the intensities into the most two distinctive groups. And that step can be performed using the student t-test with a very large degree of freedom. The test score of a threshold T applied at two sets of intensities I_1 and I_2 is defined as:

$$Score = \frac{|E[I_1]-E[I_2]|}{\sqrt{\frac{var(I_1)}{n_1}+\frac{var(I_2)}{n_2}}} \qquad (3.11)$$

I_1 and I_2 represent the observations $I_{(x,y)}$ whose averages $S_{(x,y)}$ are greater and smaller than the universal threshold, respectively, and n_1 and n_2 are the size of each group. The threshold with the highest score is chosen to be the universal threshold between the dark and the bright regions. The bias can be removed by scaling the mean of the smaller cluster to the mean of the bigger one.

The normalizing steps are summarized in the following algorithm:

Given the raw image $I_{(x,y)}$:

define outliers = $I_{(x,y)}$ s.t $I_{(x,y)} > 30 \times std(I_{(x,y)})$
eliminate the outliers

for all (x,y)
    define a window, W, centered at (x,y)
    $S_{(x,y)} = mean(I_{(x,y)} \in W)$

for T = min(S) to max(S)
    $I_1 = I_{(x,y)}$ such that $S_{(x,y)} <= T$
    $I_2 = I_{(x,y)}$ such that $S_{(x,y)} > T$
    Score(I_1,I_2) as defined in Eq.(3.11)
$T_{universal} = argMax_T(Score)$

$I_1 = I_{(x,y)}$ such that $S_{(x,y)} <= T_{universal}$
$I_2 = I_{(x,y)}$ such that $S_{(x,y)} > T_{universal}$
$n_1 = size(I_1)$, $n_2 = size(I_2)$
$\mu_1 = mean(I_1)$, $\mu_2 = mean(I_2)$

$I_{universal} = I_{raw}$
if $n_1 > n_2$
    $I_2 = \mu_1/\mu_2 * I_2$
else if $n_1 < n_2$
    $I_1 = \mu_2/\mu_1 * I_1$

We will show in the results section that, this algorithm preserves approximately 65% of the intensities and only corrects 35% of them or less.

3.5 Removal of GC-Content Bias

3.5.1 Introduction and related work

The GC content in a fragment is the percentage of the bases G and C that are included in the fragment divided on the fragment's length. GC% = (G+C)/(G+C+A+T). The GC content of a base is equal to the GC% of a 500,001bp fragment centered at the inquired base. The GC-content of 99% of the bases of the human genome is within the range from 0.34 to 0.54.

The difference between the AT and the GC content is that the GC nucleotides contain 3 hydrogen bonds whereas the AT nucleotides contain only 2. That affects the stability of the binding between the target and the probe. During the hybridization process, targets with higher GC-content are less likely to release their probe after binding to it and they are more resistant to be flushed in the washing step. Also, the extra hydrogen bond increases the fragment's florescence and the probe's intensity as well. The result of that is a strong correlation between the probe's intensity and the percentage of the GC-content in its target. It is reported in some studies [55] that the GC-content is a good predictor to estimator the raw intensity itself with a correlation coefficient of 0.994. Several studies [55],[72],[73],[74],[75] have investigated the GC content bias in the two-channel approach. The effect is observed as "waves" across the $\log_2$ ratios correlated to the GC contents [55]. Different statistical tools are used to remove the bias such as linear transformation [75], median absolute deviation [55], student t-test followed by a scaling step [73]. For more details, we refer to [75].

3.5.2 GCNORM model

Here we investigate the correlation between the GC content and probes' intensities in single channels. We present GCNORM, a new normalization model which is performed within-array not between arrays. GCNORM is a nonlinear regression model based on the GC-content percentiles, P_1, P_2, .. P_{100}. We define GC_{avg} and I_{avg} as:

$$GC_{avg}[i] = E\left[\text{GC-content}/P_{i-1} < \text{GC-content} < P_i\right]$$

$$I_{avg}[i] = E\left[I/P_{i-1} < \text{GC-content} < P_i\right]$$

Where $P_0 = 0$. The GCNORM model estimates the average intensities I_{avg} with respect to the average GC-content GC_{avg}. Figure 3.8 depicts the relationship between GC_{Perc} and I_{Perc}. Clearly, the intensity mean is directly proportional to the GC content.

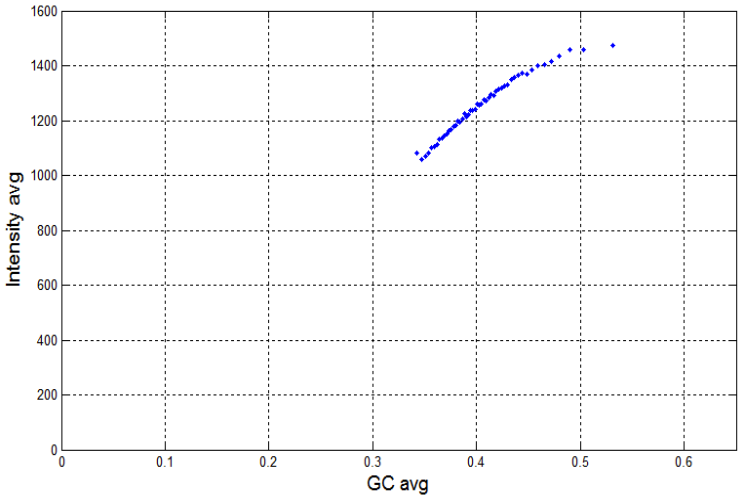


Figure 3.8: The effect of the GC-content on the intensity level.

The given range of GC-content is limited to values from 0.34 to 0.56. No measurements are given outside this interval but we can assume that the relationship is monotonically increasing and the intensity means are strictly positive. The GCNORM model defines the relationship as:

$$I_{avg} = \frac{\alpha}{1+exp[-\beta(GC_{avg}-\Delta)]} \qquad (3.12)$$

Where α, β, and Δ are real values. The last equation can rearrange as:

$$log\left(\frac{\alpha}{avg}-1\right) = -\beta(GC_{avg}-\Delta) \qquad (3.13)$$

The right-hand-side of equation (3.13) forms a straight line with respect to GC_{avg}. Therefore, α on the left-hand-side can be used as a tuning parameter to fit the LHS of the equation in a straight line with respect to GC_{avg}. To estimate the parameters for a given value of α:

$$LHS = log\left(\frac{\alpha}{I_{avg}} - 1\right)$$

$$T = \begin{bmatrix} GC_{avg} & \mathbf{1} \end{bmatrix}$$

$$\begin{bmatrix} -\hat{\beta} \\ \hat{\beta}\hat{\Delta} \end{bmatrix} = (T' * T)^{-1}(T' * LHS) \qquad (3.14)$$

$$RHS = -\hat{\beta}(GC_{Perc} - \hat{\Delta})$$

And the autocorrelation is measured as:

$$AC = \left(\frac{LHS - E[LHS]}{std(LHS)}\right).\left(\frac{RHS - E[RHS]}{std(RHS)}\right) \qquad (3.15)$$

$$\hat{\alpha} = argmax(AC)$$

Where the operator (.) is the inner product. The parameter $\hat{\alpha}$ that yields the greatest autocorrelation measurement is selected, and then $\hat{\beta}$ and $\hat{\Delta}$ are estimated using equation 3.14. The GC-content bias can be normalized using the following multiplicative operation:

$$I[i]_{normalized} = I[i]_{raw} * \left\{1 + exp\left[-\hat{\beta}\left(GC_{raw} - \hat{\Delta}\right)\right]\right\} \qquad (3.16)$$

3.6 Removal of Fragment Length Bias

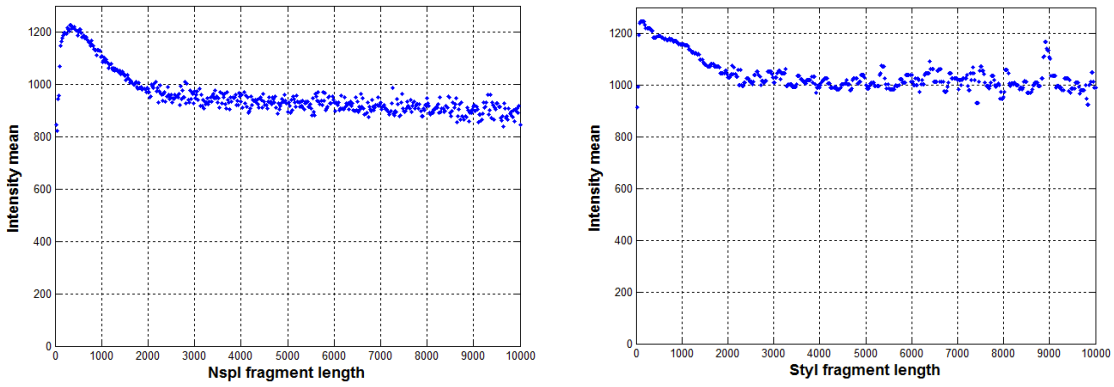
3.6.1 Introduction and related work

The bias of the fragment length on the intensity level is the most prominent and the most non-consistent source of bias in the genomic hybridization process. It occurs during the amplification process in the Polymerase Chain Reactor, PCR, which is very sensitive and prone to distortion. The PCR process is conducted in a rigorously purified

environment and under high levels of cautiousness. During the experiment, the operator is advised not to leave and re-enter the purified lab without first showering and changing into freshly laundered clothes! The accuracy of the result is very sensitive to contamination. The most likely source of contamination during the process is the residual DNA from previous experiments. A new set of instruments and proper isolation tools are strongly recommended for every new experiment.

To conduct the experiment, two identical DNA samples from the same subject are analyzed (250ng each). The two samples are fragmented using two different restriction enzymes, Sty I and Nsp I, and they are amplified separately. The result of the digestion process is short DNA fragments where 77% of them are shorter than 2000bps and 99% of them are shorter than 20,000bps. The PCR preferentially amplifies the fragments in the range 200-1000bps. The fragments are ligated to 4bps sequence that can be recognized by the PCR. The ligated fragments of each restriction enzyme are amplified separately. After that, the two amplified samples are combined and purified, denaturized, and finally hybridized to the chip. The fragments that fall within the preferred range (200-1000bps) in both channels are the fragments that are amplified the most, and thus, they have the strongest intensities.

It is widely assumed in the literature that the fragments whose lengths exceed 2000bps are not amplified in the PCR and therefore, they are ignored in the analysis. However, the PCR does amplify fragments of length up to 15,000bps. The source of the misleading information is the annotation files provided by Affymetrix. Fragments longer than 2000pbs are ignored in the annotation files and not even reported. Almost 50% of all targets fall within fragments shorter than 2000bps in both channels at the same time while the other 50% of targets fall within fragments shorter than 2000bps in one channel and longer than 2000bps in the other channel. Fragments from both channels contribute to the final reading of the intensity. That means, the fragment length bias depends on the fragment length in both channels not only on the shortest as reported by Affymetrix.



(a) (b)

Figure 3.9: Intensity mean with respect to fragment length in StyI and NspI channels

Figure 3.9 illustrates the relationship between the intensity mean and the fragment length for channel StyI and NspI. For each channel's plot, the length of the other channel's fragments was limited to the range 200-800bps to guarantee that their contribution into the intensity is optimized. In this sample, the highest intensity mean occurs at 1200bps which means that each channel contributes 600. But as the fragment length increases, the total intensity-mean remains well above the value of 600. That assures that the amplification of the PCR is not limited to fragment lengths of less than 2000bps. In this example, the amplification of fragment longer than 2000bps is 50-60% of the optimal amplification but not zero.

Several attempts have been proposed to normalize the fragment length bias [69, 73, 76, 77, 78, 79]. Linear regression [76], cubic regression [77], quadratic regression [79], and Partek Genomic Suite [80] are some approaches used to model the bias. The effect of fragments > 2000bps is not considered in any of these attempts except the latter.

3.6.2 FLNORM model

Figure 3.10 illustrates the fragment length digested by StyI and NspI restriction enzymes. The highest intensities are concentrated from 200 to 1000bps, and then from 1000 to 1400bps in both channels. Also, the figure illustrates that the intensities of fragments shorter than 2000bps are higher than longer fragments.

The 2-D model that fits the observations as illustrated in figure 3.10 must be piecewise because of the singularities of the intensities at lengths 100bps, 1000bps, 1400bps, and 2000bps. And such a model, which consists of at least 15 pieces, is tedious to determine and to handle. Therefore, we will use a non-parametric model, FLNORM, to remove the bias. The model is extracted from the observations according to their fragment lengths as:

$$I_{avg_{X,Y}}(x,y) = E\left[I(X,Y)/x-\frac{w}{2} \le X < x+\frac{w}{2}, y-\frac{w}{2} \le Y < y+\frac{w}{2}\right] \quad (3.17)$$

Where w is a length unit. The created image can be used to normalize the raw intensities according to their lengths in StyI and NspI channels.

3.7 Results and Discussion

3.7.1 The Data of Hapmap project

The Hapmap project is an international effort to map the genetic variation in the human genomes. Its latest version (III) was released in 2010 and it contains DNA copy number microarrays of 1258 individuals from 19 different populations from various

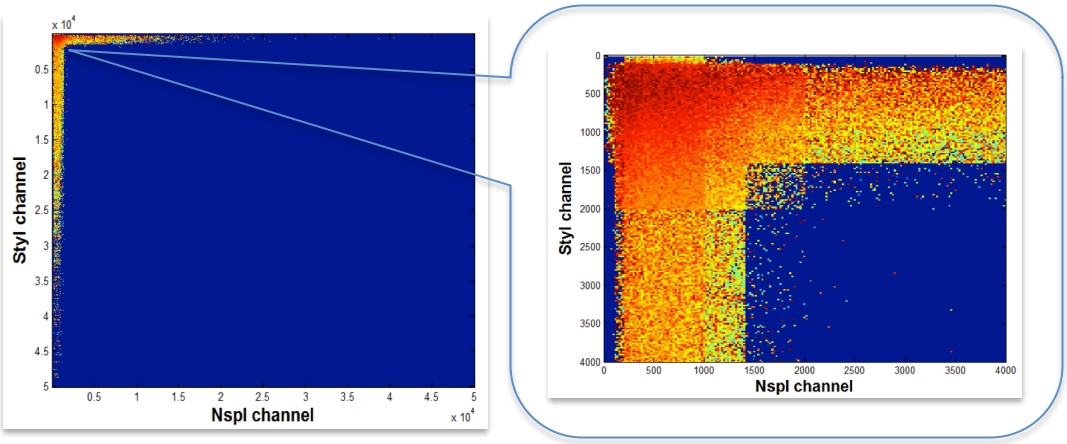


Figure 3.10: Intensity mean as a function of StyI and NspI channels.

locations on the world. For detailed information about the samples, we refer to Coriell Institute's website (http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11). The Hapmap microarrays are publicly available on the National Center for Biotechnology Information's website (www.ncbi.nlm.nih.gov). The samples were generated using Genome-Wide Human SNP Array 6.0 of Affymetrix as CEL files. We will perform the normalization methods from section 3.4-3.6 at some samples from the project. We will compare our results with the commercial software package, Partek Genomic Suite.

3.7.2 Results of the UTA algorithm

Figure 3.11 depicts the image of three randomly selected samples. A visual observation proves the existence of a spatial bias as vertical stripes of dark and bright regions. It also shows the correlation of the bias among the images. As mentioned earlier, this spatially related bias has no biological relevance and thus it is imputed to the effect of the non-ideal scanner.

The bias was localized using the edge detector in equation (3.10). A moving average window of size 21×21 was used to assign the local bias of each pixel. And the universal threshold was obtained using equation (3.11).

Figure 3.12 presents a comparison between the mean and the median moving windows in removing the bias. Because CN and SNP probes belong to different distributions, the local bias of the CN pixels is always greater than the universal threshold. Therefore, the median window is not able to normalize the bias in the CN areas. On the other side, the mean window does not discriminate between the probe groups and can normalize the bias wherever it exists.

The universal threshold segregates the observations into two groups: dark and bright. The two groups distributions are distinct which can be expected since the discontinuities

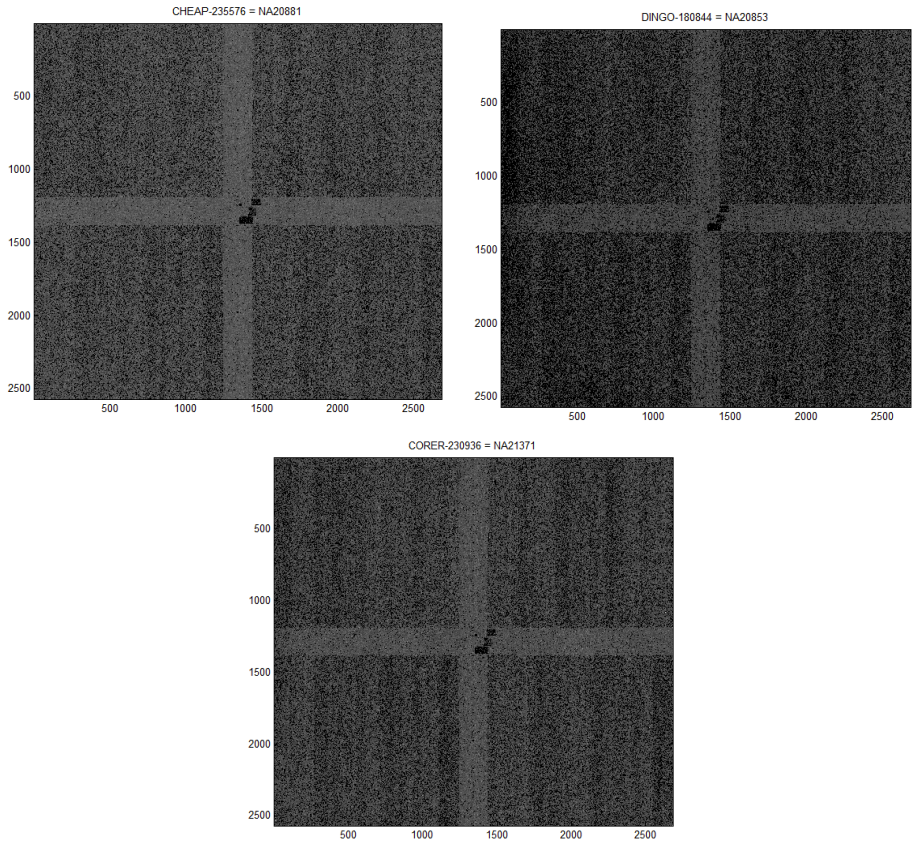


Figure 3.11: Three randomly selected samples from the Hapmap project. All samples show significant bias as dark and bright vertical stripes.

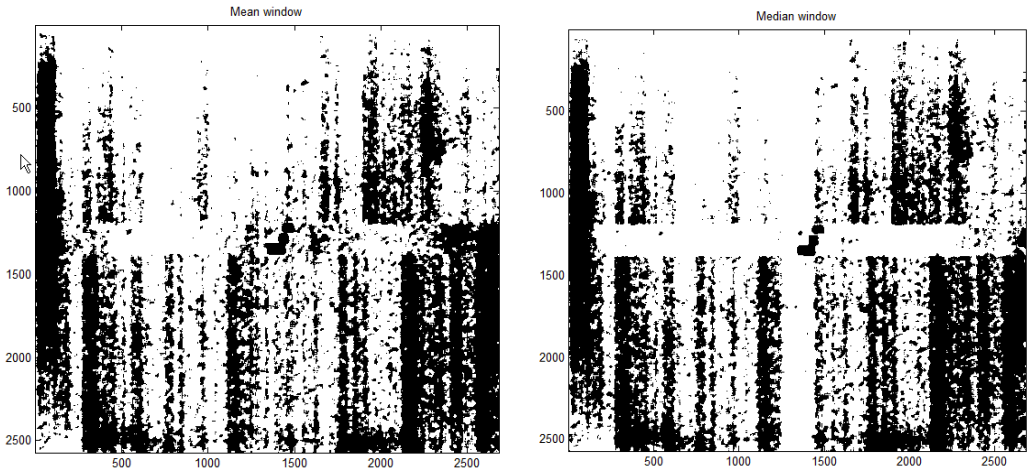


Figure 3.12: Edge detector using moving mean (left) and moving median windows (right)

between the dark and the bright regions occur abruptly. In the results: $\mu_{dark} = 952$, $\mu_{bright} = 1258$, $\sigma_{dark} = 811$, and $\sigma_{bright} = 1037$. $(\mu_{bright}/\mu_{dark}) = 1.32 \approx 1.28 = \sigma_{bright}/\sigma_{dark}$. Same result holds in all image and it is an evidence that the two distributions are just scaled from each other and thus, the bias can be removed by a basic multiplier to unify their means and standard deviations. This is further illustrated in figure 3.13 which shows the intensity distributions of the dark and bright areas in the $\log_2$ space. The multiplicative factor in the actual space is equivalent to a spatial shift in the $\log_2$ space. The two distributions are similar and the shift between them is 0.38 which is equal to $\log_2(\mu_{bright}/\mu_{dark}) = \log_2(1.3)$.

The result of UTA is presented in figure 3.14. The figure shows the image as well as the intensity means of the columns and the row before and after the normalization. Only 34% of the pixels were modified in this example. This is the main advantage of using a universal threshold rather than a local threshold where all intensities are modified.

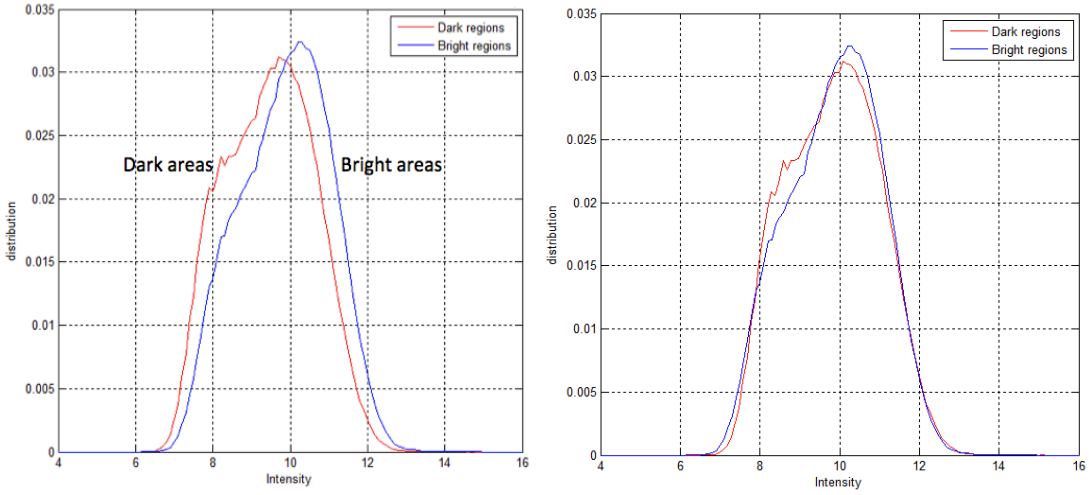


Figure 3.13: Distributions of the dark and bright areas (left) before and (right) after the normalization.

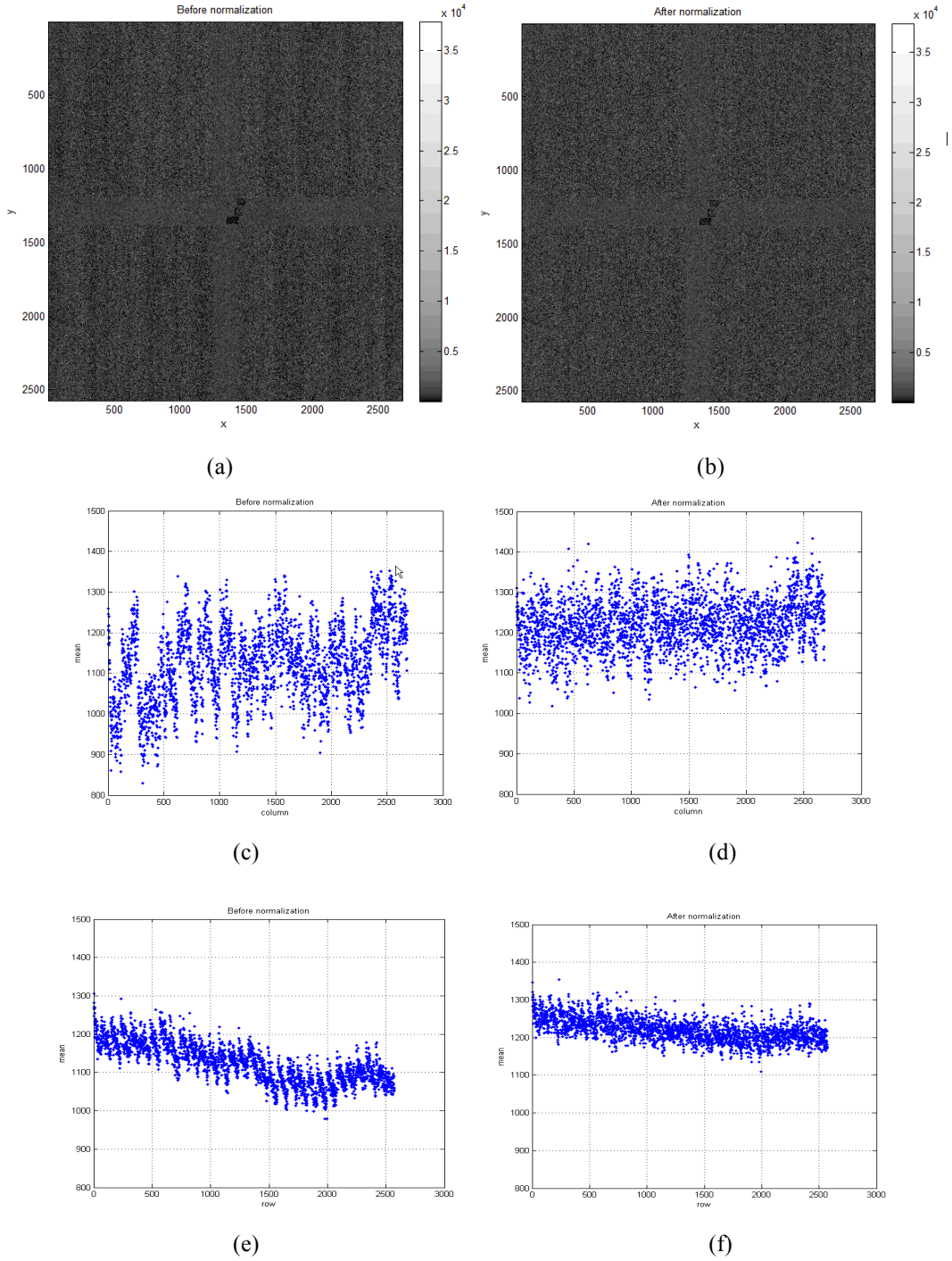


Figure 3.14: Comparison of the image's distributions before and after the normalization. (a) the original image. (b) the image after the global scaling normalization. (c) and (d) the intensity mean of the columns before and after the normalization. (e) and (f) the intensity mean of the row before and after the normalization, respectively.

3.7.3 Results of GCNORM

The parameters of the model in equation (3.12) were estimated by maximizing the autocorrelation quantity in equations (3.14) and (3.15). The model that determines the relationship between the intensity bias and the GC-content is:

$$I_{avg} = \frac{1640}{1 + exp\left[-10.5\left(GC_{avg} - 0.3\right)\right]}$$

Then, the bias can be removed by modifying the intensities as:

$$I \rightarrow I.\left\{1 + exp^{-10.5\left(GC - 0.3\right)}\right\}$$

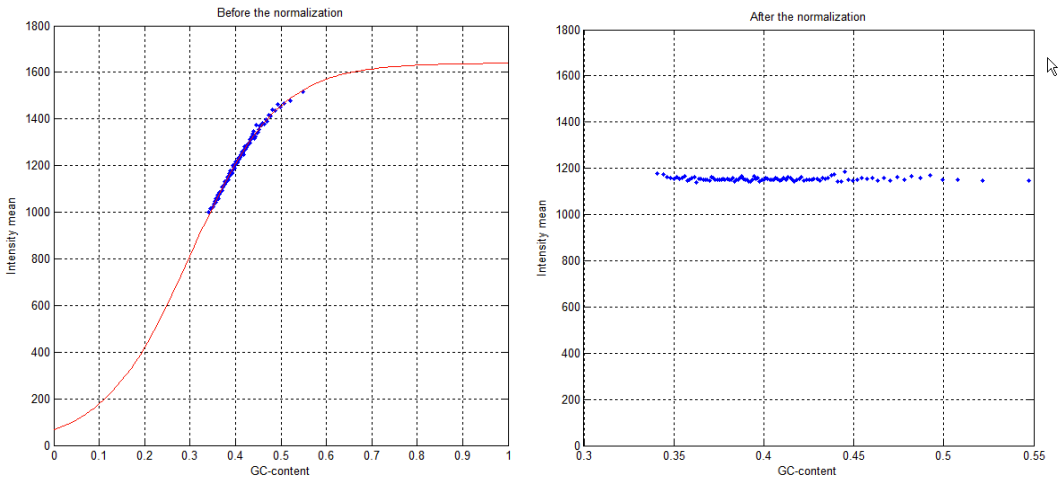


Figure 3.15: Intensity bias versus the GC-content before (left) and after (right) the normalization. The solid line is the GCNORM model of the bias.

The result of GCNORM forms a horizontal straight line as illustrated in figure 3.15. We compare the performance of GCNOMR with the normalization method embedded on Partek Genomic Suite which is a commercial package to analyze the microarrays. No details are given about the model of PGS but as illustrated in figure, 3.16, the result is not uniform. The relationship is concave downwards with its highest values centered around GC = 0.475 and it decays on both sides of its that value.

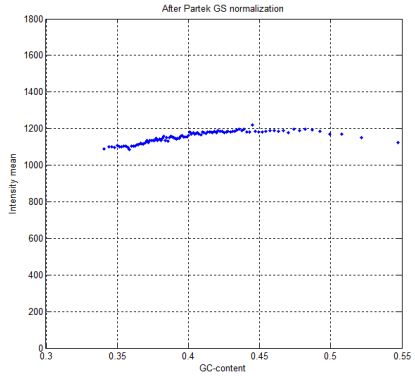


Figure 3.16: Result of GC-content normalizer in Partek Genomic Suite software.

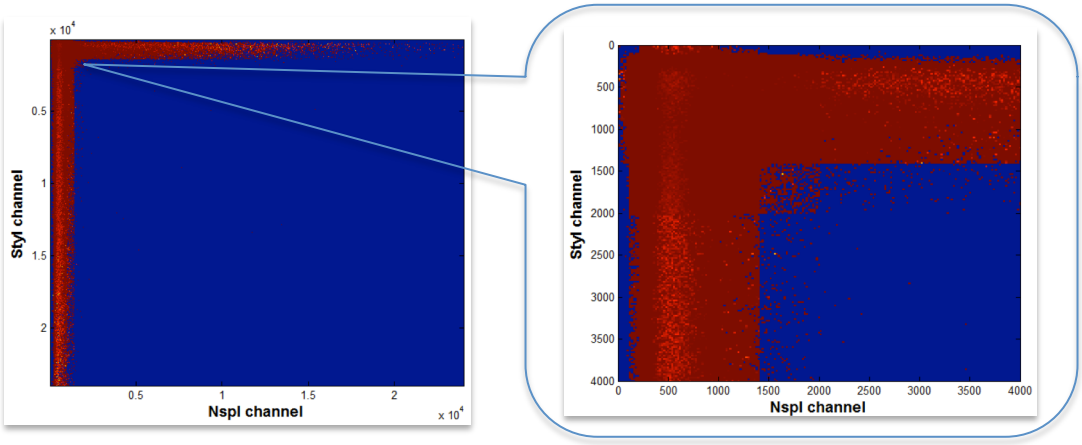


Figure 3.17: Normalized intensity mean using FLNORM.

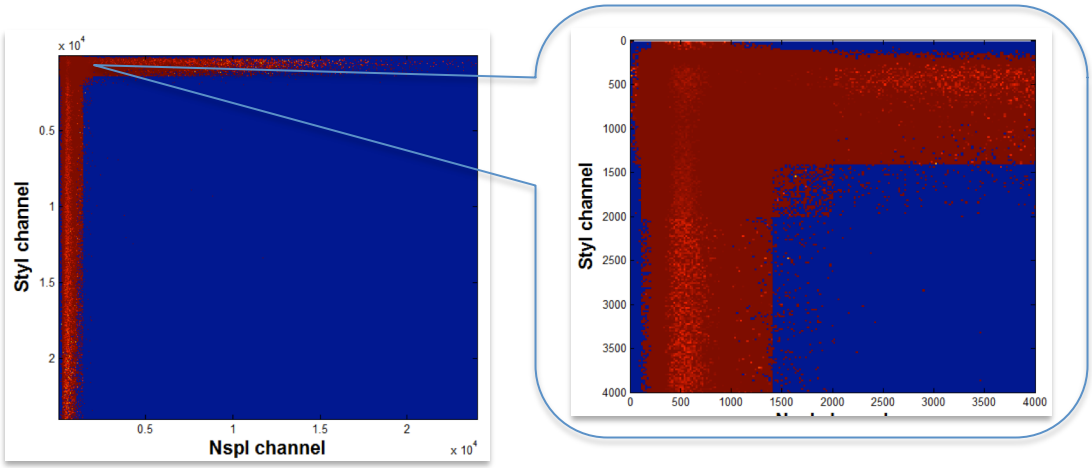


Figure 3.18: Normalized intensity mean using PGS.

3.7.4 Results of FLNORM

We choose a length unit w = 20. The model was estimated as defined in equation (3.17) and used to normalize the observations. Figures 3.17 and 3.18 illustrate the results of FLNORM and the PGS package. The results appear to be identical but we will show in section 3.8 the advantage of innovating FLNORM and its impact of reducing the computational load.

3.8 Microarrays Stationarity

We analyzed the 1258 samples of the latest release of the Hapmap project. All samples were normalized using UTA, GCNORM, and FLNORM. Also, the distributions of all samples were estimated using the QPI model. The parameters of each one of the four models converge to certain values and fluctuate around it. Figures 3.16 illustrate the ranges of the universal threshold and the scaling factor in the UTA model, the shift Δ and the rate parameter β in GCNORM, and relative mean and variance of X-chromosomes with respect to the mean and the variance of the autosomes.

From the results in figure 3.19 (a), the universal threshold in equation (3.10) is equal to:

$$\eta = 0.963\mu_{array} \quad (3.18)$$

This direct result is much cheaper to compute than searching for the optimal threshold as defined in equation (3.10). To test 100 thresholds, equation (3.11) requires 2×10^9 additions and 680×10^6 multiplications. On the other hand, equation (3.18) requires 6.9×10^6 additions and 1 division. The reduction in the computational load is very significant.

The scaling parameter to modify the dark areas of the image is 1.225.

$$I(dark) \rightarrow 1.225\times I(dark) \quad (3.19)$$

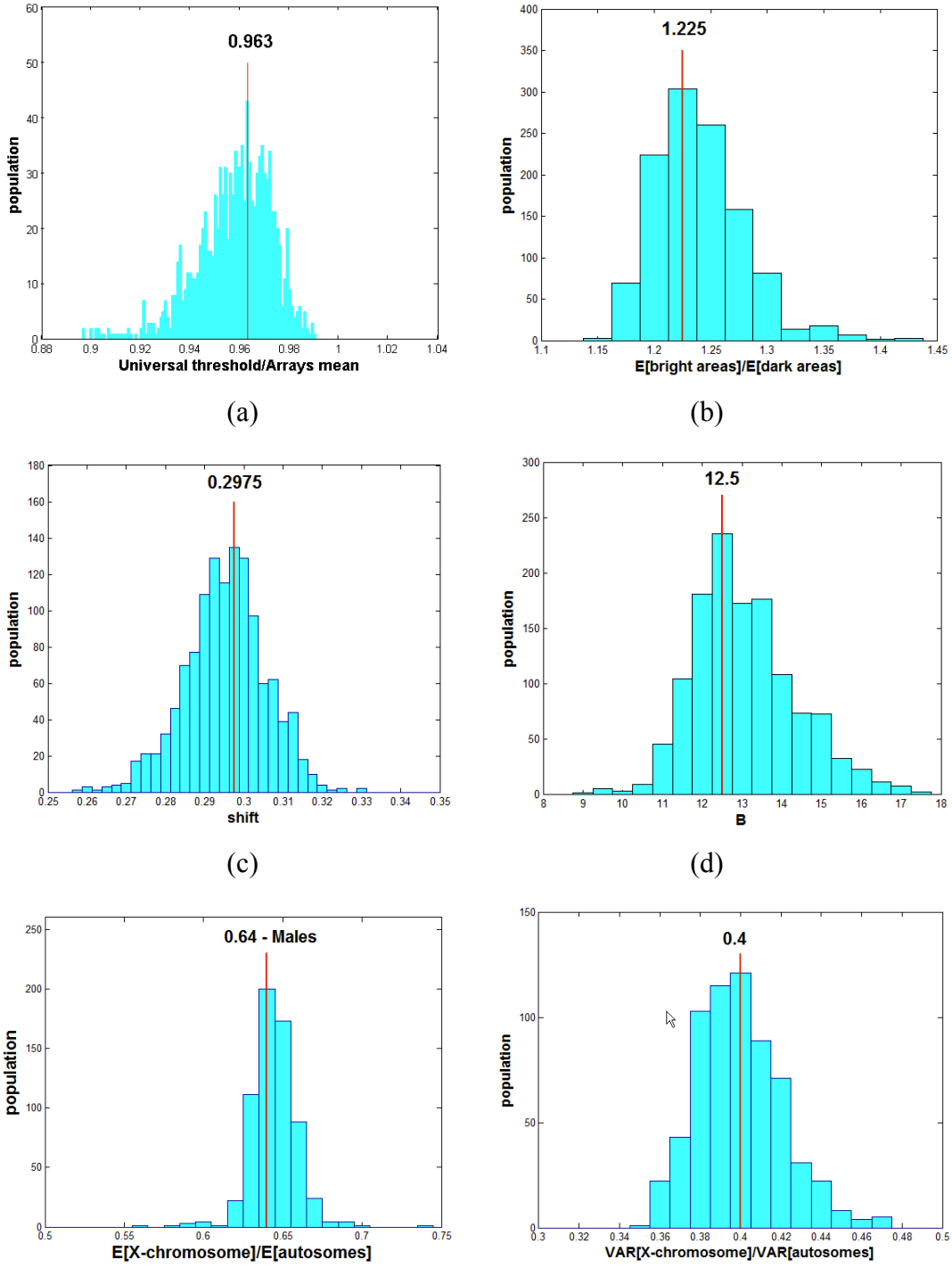


Figure 3.19: Histograms to extract (a) the universal threshold, (b) scaling factor, (c) shift Δ (d) rate parameter β, (e) relative mean, and (f) relative variance.

The GCNORM model in equation (3.12) is:

$$I_{avg} \approx \frac{\alpha}{1+exp[-12.5(\,GC-0.2975\,)]} \qquad (3.20)$$

And the bias can be removed by modifying the intensities to:

$$I[i]_{normalized} = I[i]_{raw} * \{1 + exp[-12.5(\,GC_{raw} - 0.2975\,)]\} \qquad (3.21)$$

The FLNORM model is not parametric. But we created a normalizing template which is equal to the median of the bias in the 1258 samples. The template is capable of analyzing any GWS6 sample without computing the FLNORM model for the studied sample.

Figure 3.19 (e) provides a useful relationship between the intensity mean of the X-chromosome $E[H_1]$ and the 22 autosomes $E[H_2]$. The relationship specifies that:

$$\mathrm{E}[\mathrm{H}_1] \approx 0.64\mathrm{E}[\mathrm{H}_2]$$
$$\mathrm{VAR}[\mathrm{H}_1] \approx 0.4\mathrm{VAR}[\mathrm{H}_2]$$

Which can be generalized to:

$$\mathrm{E}[\mathrm{H}_i] \approx 0.28 + 0.36\mathrm{E}[\mathrm{H}_2]$$
$$\mathrm{VAR}[\mathrm{H}_i] \approx 0.2\mathrm{VAR}[\mathrm{H}_2] + 0.2\mathrm{i}^2\mathrm{VAR}[\mathrm{H}_2] \qquad (3.22)$$

The last equation determines the intensity power of the perfect hybridization of a single DNA copy and the cross hybridization components. The means and the variances of the two components are almost equal which indicates that they have the same effect on the intensities. The intensity power of the perfect hybridization of 2 DNA copies is four times stronger than the cross hybridization.

Finally, the relationships in (3.22) can be substituted in equation (3.8) to be updated to:

$$\left.\begin{aligned} k_i &= \frac{5[0.28 + 0.36i]^2}{1 + i^2} k_2 \\ \theta_i &= \frac{1 + i^2}{1.4 + 1.8i} \theta_2 \end{aligned}\right\} \qquad (3.23)$$

For any sample, the parameters of the distribution that corresponds to the normal state k_2 and θ_2 can be estimated using equations (3.4) and (3.5). And the parameters of any other level of DCN are estimated using equations (3.23). The parameters k and θ determine the mean, mode, median, and the variance of each distribution. According to (3.23): $k_1 \approx 1.024k_2$, $k_3 \approx 0.925k_2$, $\theta_1 \approx 0.625\theta_2$, and $\theta_3 \approx 1.471\theta_2$.

3.9 Conclusions

In this chapter, we presented a novel model for the distribution of the DNA microarrays. The QPI model is robust to the outliers and to the non-homogeneity of the distribution. The model indicates that each level of the DNA copy number has its own unique distribution and all the distributions are gamma with different shape and scale parameters. We proved the stationarity of the process and the connectedness among the distributions. Knowing the quartiles of the distribution is sufficient to reveal the distribution of any level of DNA copy in the mixture.

We also presented three normalizing models: UTA, GCNORM, and FLNORM. And we showed the impact of the startionarity in reducing the computational load of the three models.

Chapter 4

Sensor Network Approach for DNA Copy Number Microarrays

Sensor network is a collection of independent sensor nodes monitoring an observation and collecting independent measurements. The network can be of any size, from 3-4 nodes to thousands or millions of nodes. The measurements of all nodes are quantized using local T-scalar quantizers, which can be identical or non-identical, and uniform or non-uniform. The input to quantizer i is a real measurement y_i and its output is a real discrete random variable $U[i] \in \{u_1,u_2,\ldots,u_M\}$ where M = T+1. See figure 4.1. All quantizers send their information to a common fusion center where it declares one of two decisions, H_0 or H_1. The efficiency of the network depends on the design of the local scalar quantizers and the performance of the fusion center. The efficiency is evaluated by calculating the average probability-of-error which depends on the network size N and the scalar quantizer's size M. This measurement becomes drastically complicated as N and M increase.

The saddle-point approximation is a powerful tool to calculate the probability of

error. Several approximations have been presented in the literature for homogenous environment. Here, we discuss the validity of the Lugannani-Rice approximation [80] in heterogeneous environments, like the DNA copy number microarrays, and we prove its accuracy using numerical results. The scalar quantizer is uniform in homogeneous environments and non-uniform in heterogeneous environments. The complexity of optimizing a uniform quantizer is linear while it is quadratic for non-uniform quantizers. We present the *Log-Lattice Lemma* to optimize the performance of the scalar quantizers in heterogeneous environments with low complexity. The saddle-point approximation and the log-lattice lemma are used to design the globally optimal fusion rule.

In chapter 3, we introduced the QPI model to estimate the distribution of the DNA microarrays. In this chapter, we introduce the sensor network approach to analyze the microarrays using the optimal fusion rule. The approach employs the results of the previous chapter and reveals the *actual* quantity of the copy number at each genomic site. The existing detection methods reveal only the variation status of duplications or deletions without quantification.

The main contributions of this chapter are: 1) a new approach of analyzing and quantifying the DNA copy number microarrays which is based on the concept of sensor networks, 2) A proof of the accuracy of the saddle-point approximation in non-homogeneous environments, 3) the *Log-Lattice Lemma* (LLL) to optimize the performance of non-uniform local quantizers, and 4) a comprehensive study of the variation in the human genome using 1258 samples of the International Hapmap Project.

In this work, we will not discuss the optimization of the network's bandwidth, capacity, energy consumption, memory, or any other physical aspects of the network. We will rather focus our attention on optimizing the performance with respect to the total error of the detection process. That includes the false positive rate (false alarm) and the false negative rate (missing) of making a decision.

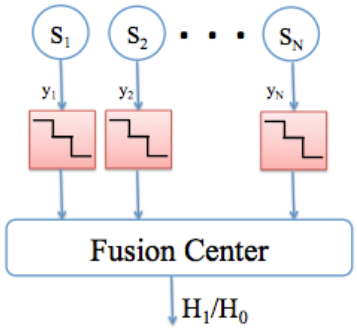


Figure 4.1: Parallel fusion network

4.1 Sensor Networks for DNA microarrays

Here we apply the sensor network model at the DNA microarrays where each probe represents a sensor node in the network. A physical network of N nodes is equivalent to a segment of the genome that consists of N probes. The analysis is performed using a moving window of size N to make a decision about the status of each genomic site, (i.e., probe). The main advantage of using the sensor network approach is that it is immune to the heavy tailed distribution of the microarray data. The heavy tails generate large amounts of outliers which have a great impact on the performance. The effect of the outliers is totally aborted by the scalar quantizers in the sensor networks. The outliers are not isolated from the rest of the observations, and therefore, they can not be eliminated using a scalar threshold.

The measurements y's belong to more than two distributions. The distributions belong to the gamma family with different shape and scale parameters. Each distribution corresponds to a different level of DNA copy number and we described how to estimate their parameters using the QPI model in chapter 3. And since the fusion rule acts in an environment of only two states H_1/H_0, the test is repeated between every two sequent events H_i/H_{i+1} to declare one of them and classify the observations into two states H_i and H_{i+1}. For example, the test can be conducted between the states H_4/H_3 to detect the genomic sites that have 4 DNA copies or more and the genomic sites that have 3 DNA

copies or less. All the duplications can be detected at once by applying the test H_3/H_2 while all the deletions can be detected by applying the test: H_2/H_1. The test needs to be repeated J-1 times to quantify the results into J levels of copy number.

Following the notations of [77], if we consider a T-scalar quantizer $\lambda = (\lambda_1, \lambda_2, \ldots \lambda_T)$ and the probability density functions of the observations, $f_1(y) = f(y/H_1)$ and $f_0(y) = f(y/H_0)$ as explained in chapter 3, the output of the quantizer takes one of M possible values under H_1 and one of M different set of values under H_0. Its probability mass functions are:

$$P_{m0} = P(\mathbf{u}/H_0) = \int_{\lambda_{m-1}}^{\lambda_m} f_0(y)dy$$

$$P_{m1} = P(\mathbf{u}/H_1) = \int_{\lambda_{m-1}}^{\lambda_m} f_1(y)dy \qquad , \text{m} = 1, 2, .., \text{T}$$

Where $\lambda_1 = -\infty$, $\lambda_M = +\infty$. The log likelihood ratio test for a network (window) of size N: $u = (u_1, u_2, \ldots, u_N)$ is:

$$\sum_{n=1}^{N} \ell_n \gtrless log\frac{\pi_0}{\pi_1} = \nu$$

Where

$$\ell_n = log\frac{\mathrm{P}(\mathbf{u}_n/H_1)}{\mathrm{P}(\mathbf{u}_n/H_0)}$$

The quantity ℓ_n is a discrete random variable and it takes one of T+1 possible values: $L_m = \log(P_{m1}/P_{m0})$. In this work, we will assume that $\pi_1 = \pi_0 = 0.5$, which means the fusion threshold $\nu = 0$.

The optimal performance corresponds to the minimum average probability of error $P_e = (P_{e0} + P_{e1})/2$. P_{e0} is the probability of the false alarm = $P(H_1/H_0)$ whereas P_{e1} is the probability of missing = $P(H_0/H_1)$. According to [77], P_{ei} can be computed as:

$$P_{ei} = \sum_{N_1,N_2,\ldots,N_M} \left[\prod_{m=1}^{M} (P_{mi})^{N_m}\right]\left[\prod_{m=1}^{M} \binom{N - \sum_{k=1}^{m-1} N_k}{N_m}\right]$$

Such that $\sum_{m=1}^{M} N_m = N$ and $\sum_{m=1}^{M} N_m L_m \gtrless 0$ (4.1)

The last function is the exact formula of computing the probability of error and it contains approximately N^{M-1} terms. This is not impossible to compute if the optimal quantizer is given and it consists of 4 or less. But the problem gets tedious as the quantizer's size increases. Furthermore, if the variances of H_1 and H_0 are not equal (heterogeneous), which is the case in all microarrays, then the optimal quantizer is non-uniform and it requires approximately N^{M-1} operations to compute. That makes the total complexity of minimizing the average error in the order of N^{2M-2}, which makes the optimization problem infeasible. As an example, the complexity to minimize the error for a small network of 25 nodes with a 4-ary quantizer is 244×10^6 and 3.7×10^{19} with an 8-ary quantizer. To overcome this obstacle, we present two methods: the Lugannani-Rice formula of the saddle-point approximation to reduce the complexity of computing the average error, and the Log-Lattice Lemma to reduce the complexity of finding the optimal T-scalar quantizer.

4.2 Saddle-point approximation

The saddle-point is a popular method to compute the tail probability of the sum of independent and identically distributed random variables. Many approximation of the exact formula have been derived using the cumulative generic function and its derivatives. The Lugannani-Rice approximation is one of the easiest and most efficient approximations [80]. It is based only on the first and the second derivatives of the cumulant generic function. The moment generating function of ℓ_n is:

$$G(\theta) = \sum_{m=1}^{M} P_{m0} e^{\theta L_m} = \sum_{m=1}^{M} P_{m1} e^{\theta L_m}$$

And the cumulant generating function and its first and second derivatives are:

$$\begin{aligned} K(\theta) &= log G(\theta) \\ K'(\theta) &= \partial K(\theta)/\partial\theta \;=\; W_1(\theta)/G(\theta) \\ K''(\theta) &= \partial^2 K(\theta)/\partial\theta^2 \;=\; [G(\theta)W_2(\theta) - W_1(\theta)^2]/G(\theta) \end{aligned} \qquad (4.2)$$

Where

$$W_k(\theta) = \sum_{m=1}^{M} P_{mi} (L_m)^k \, e^{\theta L_m}$$

The saddle point θ, is obtained by solving the equation:

$$K'(\theta) = \; W_1(\theta)/G(\theta) = \upsilon/N = 0$$

Since we assumed $\pi_1 = \pi_0 = 0.5$. The last equation implies that:

$$W_k(\theta) = \sum_{m=1}^{M} P_{mi}. L_m. e^{\theta L_m} = 0$$

This equation can be solved using any method of the household's methods. By solving the equation and obtaining the saddle point θ, K(θ) and K″(θ) can be obtained directly from equation (4.2). And then:

$$\begin{aligned} P_{e0} &= \Phi(\mathrm{r}) + \varphi(r)[q^{-1} - r^{-1}] \\ P_{e1} &= \Phi(\mathrm{r}) - \varphi(r)[q^{-1} - r^{-1}] \end{aligned} \qquad (4.3)$$

Where

$$r = Sgn(\hat{\theta})\sqrt{-2NK(\hat{\theta})} \qquad (4.4)$$

$$q = \hat{\theta}\sqrt{NK''(\hat{\theta})} \qquad (4.5)$$

The Lugannani-Rice approximation in [80] is well established to compute the average probability of error for the sum of continuous random variables. It is proved in [77] that the approximation is accurate for the discrete random variables in homogenous environments. And here we prove its accuracy for the sum of discrete random variables in non-homogenous environments. We will present numerical results in section 4.4 but first we need to address the problem of optimizing the local quantizer.

4.3 Log-lattice quantizer

The process of finding the optimal design of the T-scalar quantizer in microarrays environments is substantially different than the physical sensor networks. The physical sensor networks are subject to identical additive noise while the microarrays are subject to additive and multiplicative noise. The only effect to H_1 and H_0 under the additive noise is a shift in the mean value while the variance is identical under the two states. If the noise is multiplicative, both the mean and the variance are affected, and that has a remarkable impact on finding the optimal T-scalar.

In the case of identical variances, the scalar quantizer is lattice and its middle term is equal to the mid-point between μ_1 and μ_0 where $\mu_i = E[H_i]$. A variable is lattice when the difference between any two of its terms is equal to $n\beta$ where β is the lattice span β and n is an integer. The most basic representation of a uniform lattice scalar is:

$$\boldsymbol{\lambda} = \left(\dots, \ \frac{(\mu_1 + \mu_0)}{2} - 2\beta, \ \frac{(\mu_1 + \mu_0)}{2} - \beta, \ \frac{(\mu_1 + \mu_0)}{2}, \ \frac{(\mu_1 + \mu_0)}{2} + \beta, \ \frac{(\mu_1 + \mu_0)}{2} + 2\beta, \dots \right)$$

This quantizer is uni-variate and it depends only on β. The optimal quantizer can be found by applying an exhaustive search over the space of β. The complexity of the process is in the order of O(n). Combining this step with the Lugannani-Rice approximation makes the problem of minimizing the fusion rule's average error in the order of $O(n^2)$.

If the variances are not equal, then the optimal quantizer is not lattice. The problem of finding the optimal T-scalar quantizer is multivariate in T variables and its complexity is in the order of $O(n^T)$, which is extremely tedious for T > 4. As a numerical example, if T = 7 (8-ary quantizer) and the quantizer's domain is divided into 100-point grid, then finding the optimal quantizers requires computing the saddle-point approximation $\binom{100}{7} \sim 1.6*10^{10}$ times. This is infeasible and the solution must be acquired in a more applicable approach. Here we present the log-lattice lemma to solve the problem of obtaining the optimal non-uniform quantizer efficiently.

Log-lattice Lemma (LLL)

If the variances of H_1 and H_0 are not equal ($\sigma_1^2 \neq \sigma_0^2$) and the log likelihood function $\ell_n = \log\frac{P(u_n/H_1)}{P(u_n/H_0)}$ is monotonic, then the following T-scalar quantizer converges asymptotically to the optimal quantizer as $T \rightarrow \infty$.

$$\lambda_i = \tilde{\mu}\beta^{\left\{i-\left(\frac{T+1}{2}\right)\right\}}, \; i = 1,2,\ldots \qquad (4.6)$$

Where $\tilde{\mu}$ is defined as the point where $P(y/H_1) = P(y/H_0)$:

$$P(Y = \tilde{\mu}/H_1) = P(Y = \tilde{\mu}/H_0)$$

Another form of the log-lattice quantizer is:

$$\boldsymbol{\lambda} = (\ldots, \; \tilde{\mu}\beta^{-2}, \; \tilde{\mu}\beta^{-1}, \; \tilde{\mu}, \; \tilde{\mu}\beta^{+1}, \; \tilde{\mu}\beta^{+2}, \ldots)$$

We called it *log-lattice quantizer* because the vector log($\boldsymbol{\lambda}$) per se is lattice. The problem of finding the optimal quantizer using the log-lattice lemma can be solved using an exhaustive search over the span of β, and its complexity is reduced to O(n).

4.4 The accuracy of the saddle-point approximation in non-homogenous mixtures

The accuracy of the saddle-point approximation in sensor network is well proved and discussed in many literatures. However, the work has been limited to the homogenous case where the difference between the distributions of H_1 and H_0 is just a spatial displacement. Here we present numerical results to prove the accuracy of the saddle-point approximation for the non-homogenous mixtures. Later, we will employ the saddle-point approximation in detecting the variation in the DNA copy number arrays in one-channel and two-channel approaches.

We ran three experiments:

Exp1: $H_0 \sim \Gamma(3,1000)$, $H_1 \sim \Gamma(4,1500)$ $\Rightarrow$ SNR = -4.3dB with $\boldsymbol{\lambda} = (-\infty, 1994, 3390, 5763, \infty)$

Exp2: $H_0 \sim \Gamma(3,1000)$, $H_1 \sim \Gamma(4,1250)$ $\Rightarrow$ SNR = -6.7dB with $\boldsymbol{\lambda} = (-\infty, 2106, 3580, 6086, \infty)$

Exp3: $H_0 \sim \Gamma(3,1000)$, $H_1 \sim \Gamma(3.5,1250)$ $\Rightarrow$ SNR = -9.5dB with $\boldsymbol{\lambda} = (-\infty, 2353, 4000, 6800, \infty)$

In all cases, $\pi_1 = \pi_0 = 0.5$. We adopt the following definition of the signal-to-noise ratio for the non-homogenous mixtures of gamma distributions:

$$SNR = 10.log_{10}\left\{ \frac{(k_1\theta_1 - k_2\theta_2)^2}{2(k_1\theta_1^2 + k_2\theta_2^2)} \right\} \text{dB}$$

We used the formula in (4.1) to measure the exact average probability of error. The difference between the saddle-point approximation and the exact formula in measuring the average probability of error is presented in figure 4.2. Both values decay asymptotically as well as the difference between them. However, the relative error ($100\times|P_{exact} - P_{saddle}|/P_{exact}$) reaches its steady state for N>30 and at that range $\lim_{N\to\infty}\{P_{saddle}/P_{exact}\} \approx (1+\varepsilon)$ where ε is a small positive real number. The approximated average probability error converges asymptotically to the exact value in the homogenous environments [43] but not in the heterogeneous environments.

It is interesting to notice that the saddle-point approximation in the experiment with the lowest SNR is the closest to the exact value of the probability of error. The relative error is 2% for SNR = -9.5dB compared to 10% for SNR = -4.3dB. Although the measurement of the approximated value is not 100% accurate, it is still practical for use for two reasons. First, the approximated value is still convergent to the exact value and the difference between the two measurements is decaying asymptotically. And second, the approximated value is much more efficient in the computational load. Our machine spent about 5 minutes to generate the exact value of the average probability of error for networks of N sensor where N = 1, 2, 3, …, 70. The same machine spent about 1 second to generate the approximated results. The number of operations required to obtain the approximated value is 3M+7, and that amount is required for each iteration of the household's method. And since the method requires usually 10 iterations, that makes the total number of operations less than 200. It is much less than 70^3 = 343,000 operations required by the exact formula.

4.5 The accuracy of the log-lattice lemma

We will employ the measurement of the asymptotic relative efficiency (ARE) to evaluate the accuracy of the LLL in equation (4.6). The ARE is a useful measurement to compare the results of using the scalar quantizer with the results of using the real un-quantized values. It is defined as:

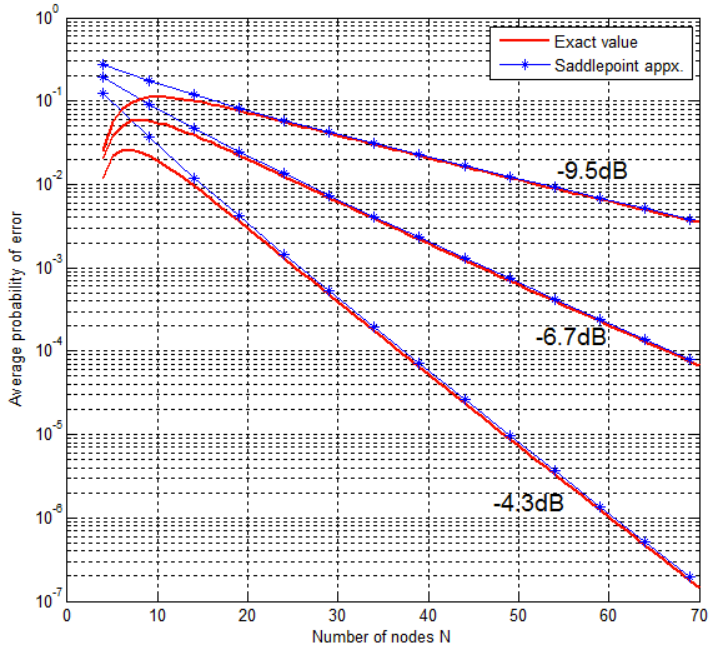


Figure 4.2: The average probability of error versus the number of nodes.

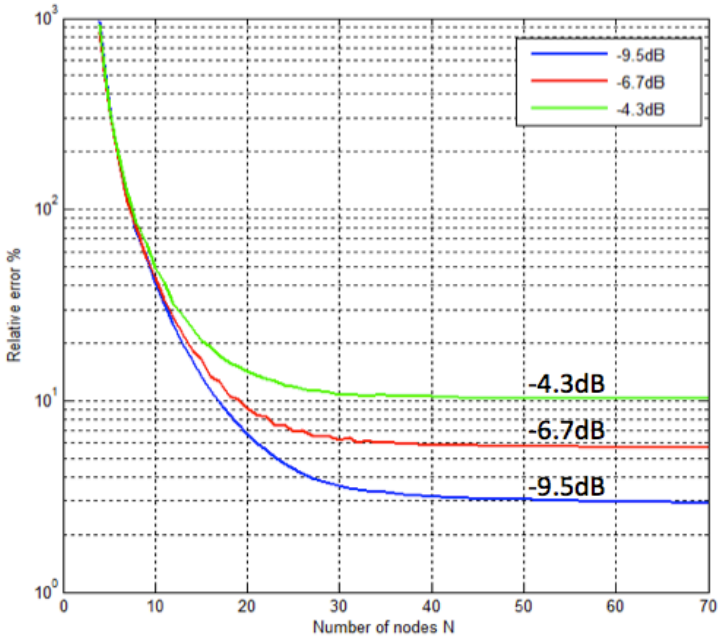


Figure 4.3: Relative error of the P_e versus the number of nodes.

$$ARE = \frac{C_M}{C_\infty}$$

C_M is always less than C_∞ and they are defined as:

$$C_\infty = \ argmax_{0\le a\le 1} \ -log\left\{\int_{-\infty}^{\infty} f(y/H_1)^a . f(y/H_0)^{(1-a)}\, dy\right\}$$

$$C_M = \ argmax_{0\le a\le 1} \ -log\left\{\sum_{m=1}^{M} (P_{m1})^a . (P_{m0})^{1-a}\right\}$$

As the number of the quantizer's tabs increases, the value of the ARE converges to $C_\infty$and the accuracy improves. ARE → 1 as M → ∞. We calculated the ARE values for the three experiments mentioned in section 4.4 using a network of 25 sensors, N = 25. We changed the size of the scalar quantizers gradually from 4-ary quantizer to 50-ary. The results are illustrated in figure (4.4). The ARE curves are identical in all experiments regardless the fact that they have different levels of SNR. The value of the ARE exceeds 0.995 with M = 32 and exceeds 0.987 with M = 16. That means the loss of the quantizers is less 1.5% using a quantizer of 16 tabs or more. It also proves that the log-lattice quantizer is optimal.

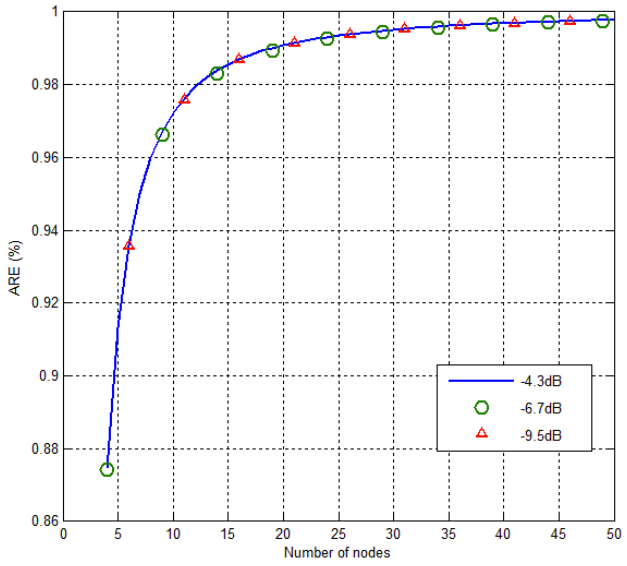


Figure 4.4: The asymptotic relative efficiency versus the number of nodes.

4.6 Experimental results of the sensor networks approach

4.6.1 Two-channel approach

We tested the same data that we used in our comprehensive analysis in section 2.5. We used the sensor network approach with the saddle-point approximation as described in section 4.2 since the two-channel approaches are assumed to be homogeneous. The sensitivity was measured by comparing the results of the sensor networks approach with the results of the QPCR while the false alarm was measured using the self-self arrays.

We normalized the 7 arrays to have $\sigma^2 = 1$. The performed test is:

$$H_0 \sim N(0,1) \quad \text{and} \quad H_1 \sim N(u,1).$$

Where u ranges from 0 to $+\infty$ to detect the duplications and it ranges from 0 to $-\infty$ to detect the deletions. We used quadratic quantizer, T=3 and M=T+1 = 4. The optimal quantizer for each u is computed using the saddle-point approximation. The quantizer is centered at +u to detect the duplication and centered at –u to detect the deletions, and hence, the process is done twice. After mapping the observations using the optimal quantizer, a moving average window of size 25 is applied. The moving average window is an analogy of the actual sensor network which consists of 25 sensors. The ROC curve of the results is shown in figure (4.5). The figure also shows the performance of TLRT and MIS. We showed in section 2.5 that the TLRT and the MIS outperform 26 different methods available in the publications and software packages. And here we compare the performance of the sensor networks approach with the two methods.

The optimal quantizer is $\lambda = (-\infty, -1.38, 0, 1.38, +\infty)$ and $u = 1$. The likelihood vector is: (-3.706, -1.183, +1.183, +3.706) as shown in figure (4.6). The performance of the sensor network approach, which is based on the likelihood test, outperforms the TLRT at $P_f < 5\%$. It almost matches the curve of the MIS method where the false positive rate at the edges decays $\sqrt{N}$ times faster than the likelihood tests.

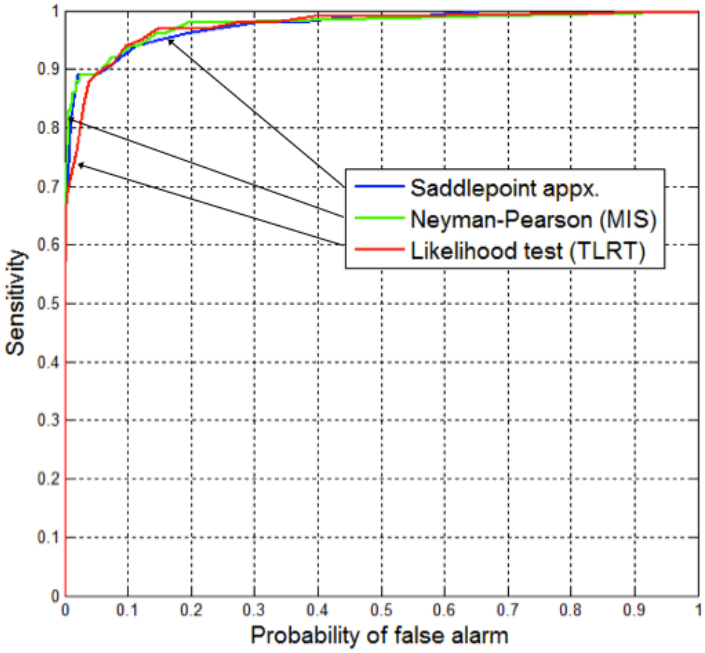


Figure 4.5: ROC curves of TLRT, MIS, and the saddle-point approximation

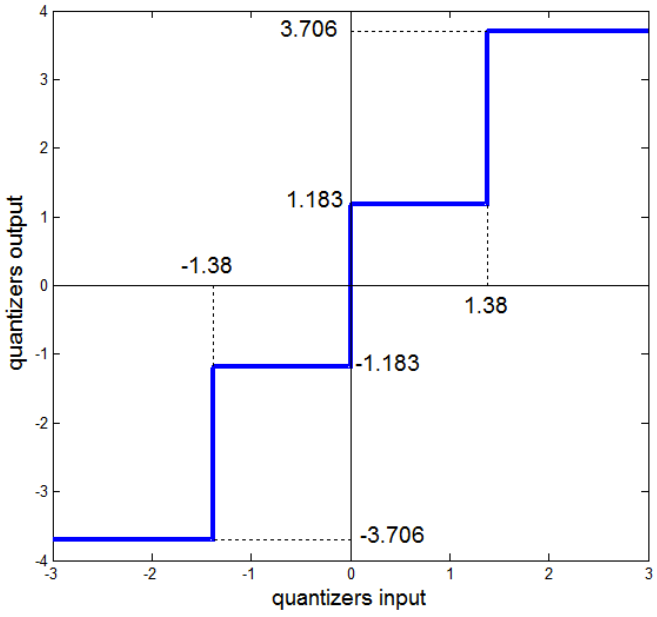


Figure 4.6: Optimal quantizer to detect the CNV in the two-channel microarrays

The model in figure (4.7) is capable of analyzing any sample of the two-channel microarrays.

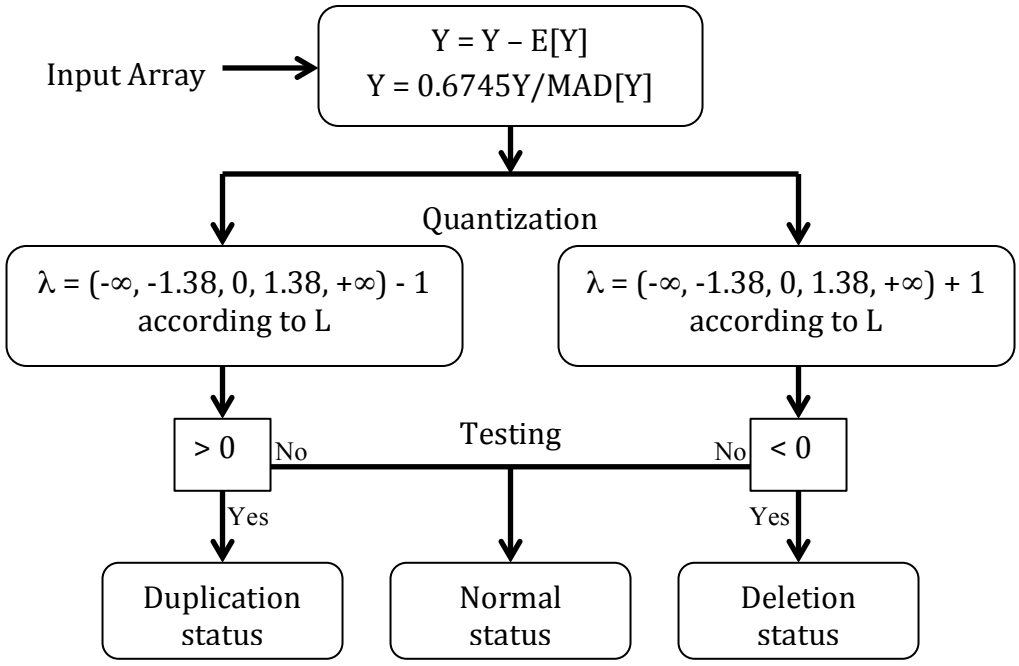


Figure 4.7: Optimized model to analyze two-channel arrays

In the ideal case, the average probability of error of Gaussian distributions with distance = 2 between their mean is equal to 3.66×10^{-6}, well smaller than the values we are getting in the results. That proves the model of the observations is not perfectly Gaussian, and another model could be better in fitting the data. It also proves that, the outliers are not isolated or disconnected from the rest of the observations. The optimal quantizer indicates that, the outliers effect at least the top 15% of each side of the distribution, and that requires more awareness during the analysis.

4.6.2 The effect of the number of molecules and the network size

We ran several experiments using simulated data to study the effect of several parameters. The distributions of the test's states are:

$H_1 \sim \Gamma(0.925k,588)$, $H_0 \sim \Gamma(k,400)$, to detect the duplications

$H_1 \sim \Gamma(1.024k,250)$, $H_0 \sim \Gamma(k,400)$, to detect the deletions

The data consists of 200 segments of duplication separated by non-variant segments. Each segment of duplication has a unique length ranging from 26 to 225 points. The separating normal segments have a common length of 500 points. There are also 200 segments of deletion with lengths from 26 to 225 and separated by normal segments of length 500 points. The test was repeated for several values of k = 2, 2.5, 3, 3.5, and 4. For each case, the network size N was changed from 5 to 301. We applied the sensor network approach on the simulated data and we measured the sensitivity and the false alarm for each size of the network. The sensitivity = $P(H=H_1/H_1)$ while the false alarm = $P(H=H_1/H_0)$. The results are shown in figure 4.8.

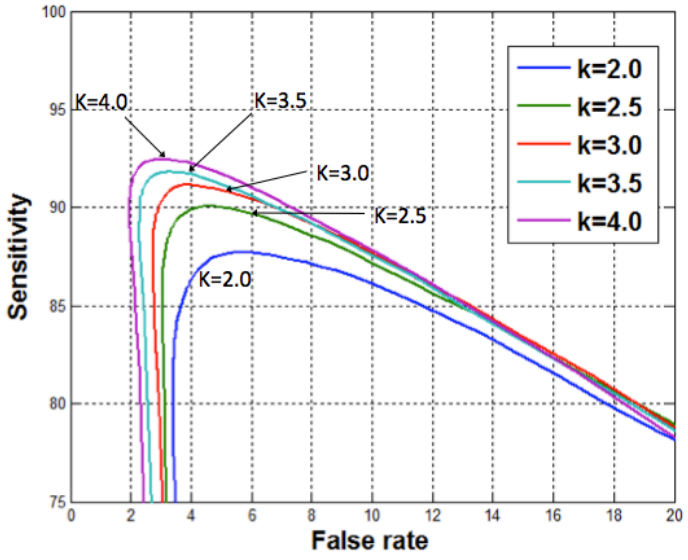


Figure 4.8: ROC curves for several values of shape parameter and network size

The results show that, the performance is optimal when the network size is within the range from 45 to 65. They also show that the performance gets better as the shape parameter k increases. The shape parameter depends on the intensity mean which depends on the total number of targets in the assay which includes the targets from n molecules of DNA. The conclusion is that, it is preferred to increase the shape parameter to optimize the performance. That can be done by using bigger samples than the 250 ng of DNA specified by Affymetrix or by adding more amplifying cycles to the PCR.

4.6.3 The stability and variability of the human genome

We analyzed the 1258 samples of the International Hapmap Project in its third release. The samples are analyzed using Affymetrix GWS6 arrays and they are publicly available at: http://hapmap.ncbi.nlm.nih.gov/downloads/raw_data/hapmap3_affy6.0/. The main goal of our analysis is to assess the stability and variability of the human genome. We will not consider the differences in gender, age, race, global location, or any other discriminative factors in this work. We will only consider the common variations in the whole genome or in a specific chromosome.

All samples were normalized using the UTA, GCNORM, and FLNORM models as described in section 3.4, 3.5 and 3.6. Then the distribution of the result was estimated using the QPI model as described in section 3.3. Then the variation was detected using the sensor network approach as described in sections 4.2 and 4.3. The optimal size of the network was obtained using the simulated experiment in section 4.6.2. The false alarm rate is within the range from 2% to 5%.

The novelty of the sensor network approach is not only its high accuracy and low complexity, but it extends to the type of its results. While the existing methods detect only the variation of the copy number, the sensor network approach detects *and quantifies* the level of variation. The importance of quantifying the variation emanates from the direct effect of the copy number amount on the gene's dosage [2]. A gain of

one copy might increase the gene's production by up to 50% while a gain of 3 copies might increase the production by up to 150% and the difference between the two cases is very significant. We found in our results that, the detected copy number ranges from 0 to 6. i.e., two-copy loss to four-copy gain. The high-copy-repeat sequences are usually avoided in the hybridization experiments and not considered in chip's design.

Figure 4.9 illustrates the rate of total duplication and total deletion according to their frequencies in the samples. There different criteria to express the results according to the frequency of the variation. We will adopt 4 criteria: 25%, 50%, 75%, and 95%. 25% means that the variation exists at least in 25% of the tested samples. Considering the 50% criteria, the results indicate that 11.7% of the human genome is not diploid. That includes 2.5% of deletion and 9.2% of duplication common at least in 50% of the samples. This percentage is significantly larger than the percentage reported in [1] where 12% of the genome is not diploid using less than 1% criterion. The results also indicate that about 30% of the genome is stable in all samples.

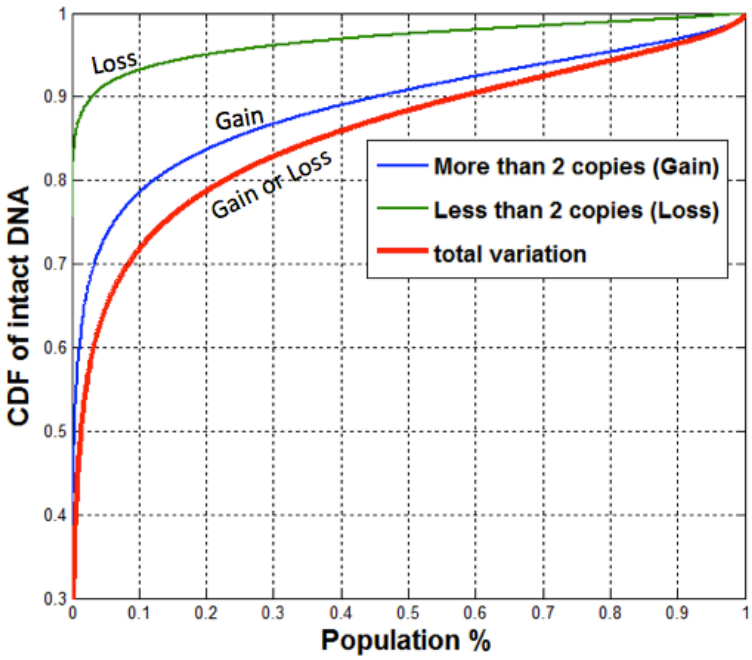


Figure 4.9: Duplication and deletion rates in normal human genome.

Figure 4.10 illustrates the population of different DNA copy number quantities in the samples. The total sum of these variations accounts for about 1-2% of the genome using any criterion. As expected, the frequency of the duplication decreases as the variation level increases. Using the 50% criterion, about 0.02% of the genome has 4 DNA copies and almost zero percentage of the genome has a copy number above 4 or less than 1.

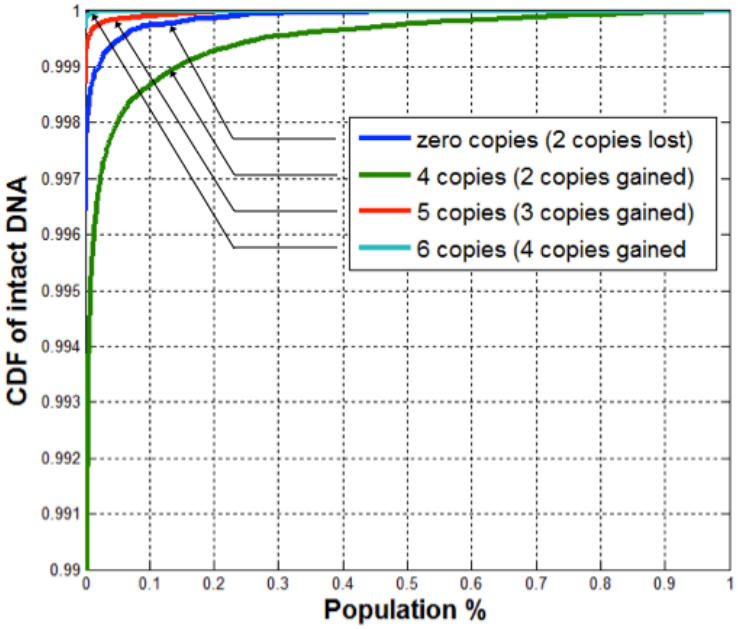


Figure 4.10: Quantified duplication and deletion rates in normal human genome.

The frequency of duplications and deletions based on the chromosome are presented in in figure 4.11 using three criteria 25%, 50%, and 75%. In general, the duplication rate is about 3 times the deletion rate in each chromosome and over the whole genome. Chromosomes 16, 17, 19, and 20 exhibit the highest rate of duplication while chromosomes 4, 13, and 18 exhibit the lowest rate of duplication. There is a visible pattern between the duplications and the deletions on chromosomes level at 25% criterion. Each chromosome has a tendency to have more duplication or more deletion, but not both. The pattern gradually decays until it disappears at 95% criterion as shown in figure 4.12.

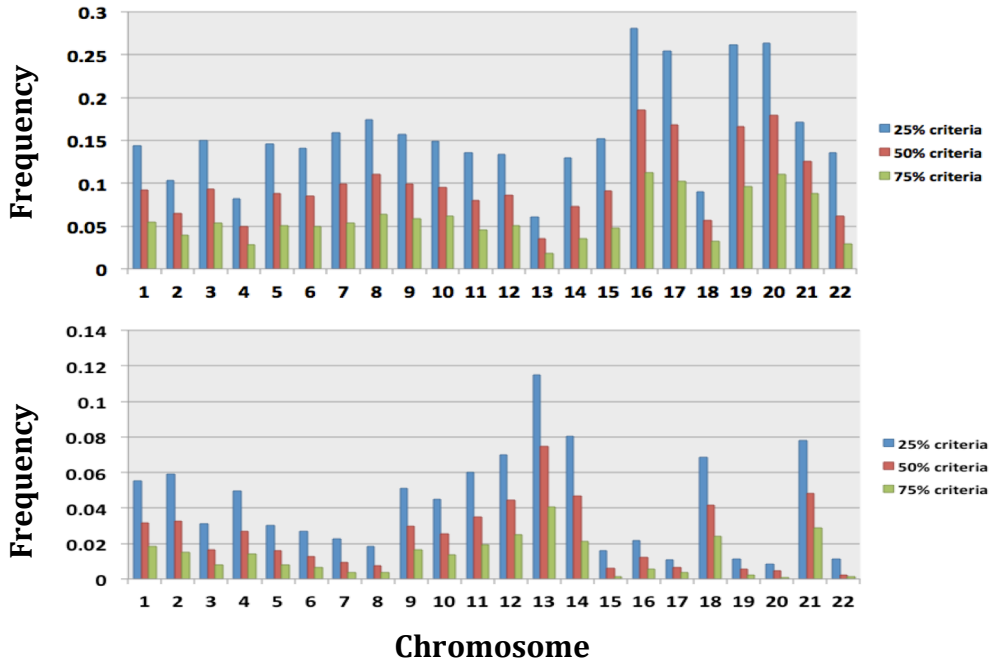


Figure 4.11: Chromosomal duplications (up) and deletion (down) rates using 25% (left bar) 50% (middle bar) and 75% criteria (right bar)

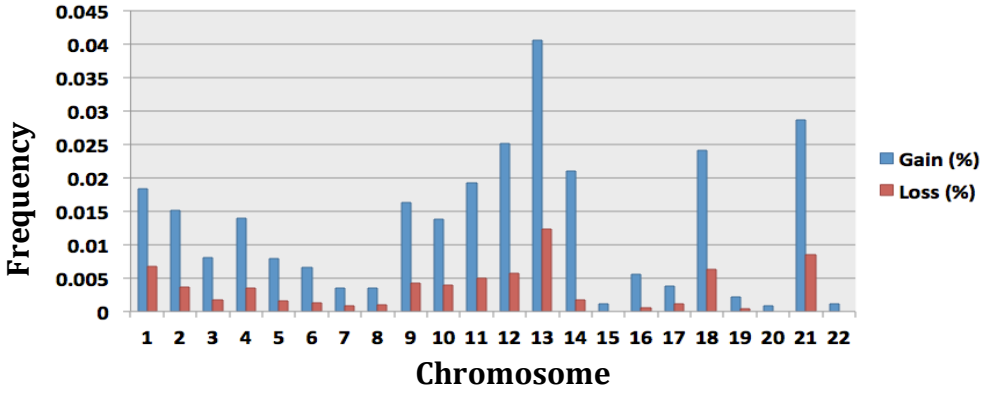


Figure 4.12: Chromosomal duplications (long bars) and deletion (short bars) rates using 95% criterion

Finally, the stability of each chromosome is measured using 1% criterion. In other words, the stability is measured as the percentage of the chromosome that is diploid in at least 99% of the population. Slightly more than 50% of chromosomes 4, 13, and 18 are stable while the stability is around 35% of chromosomes 16, 17, 19, and 20. Other chromosomes have a stability between 40% and 50% as shown in figure 4.13.

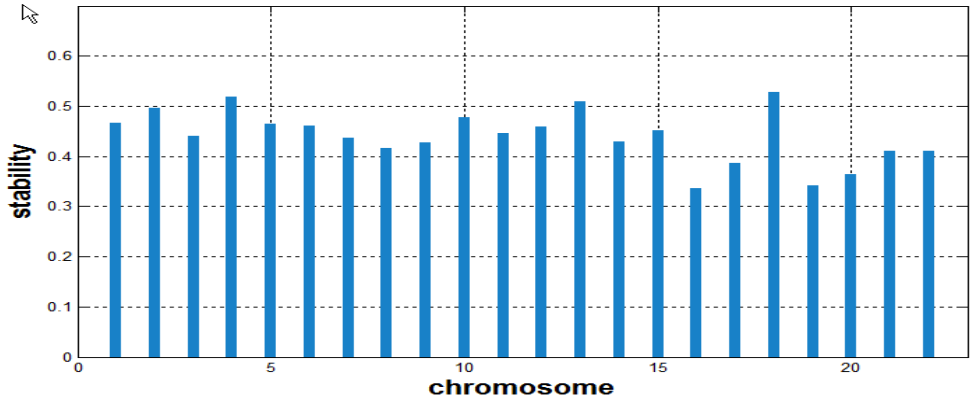


Figure 4.13: Chromosomes stability using 1% criteria

4.7 Conclusions

We proved the accuracy of the saddle-point approximation and the log-lattice lemma for heterogeneous environments. The two techniques were employed in the sensor network approach to analyze single-channel DNA microarrays with high accuracy and low computational load.

The noise in the microarrays is multiplicative and the signal-to-noise ratio increases as the total number of targets increases. It is recommended to use bigger samples of DNA or apply more amplification cycles of the PCR to boost the accuracy.

The stability and variability of the human genome is not fully evaluated yet. We presented results of the copy number variation in 1258 samples using different criteria. We showed that about 40% of the genome is diploid in all people while 11.7% is variant in at least 50% of all people.

Chapter 5

Correlation Between Copy Number Variation and Human Diseases

The methods in the previous chapters provide different tools to detect the variation the alteration of DNA copy number but they don't explore the connection of the variation from one result to another. The question is: is there a correlation or certain patterns of DNA copy number variation that can lead to a better understanding of the genome's functions and can be used as a biomarker to detect a disease or the person's susceptibility to it? We will explore several clustering algorithms to find the answer.

The main contributions of the this chapter are 1) Segregation-Based Subspace-Clustering algorithm SBC to reveal patterns of similarity among special objects, and 2) we present results of two experiments to show the correlation of the DCV with autism and the activity of the androgen depletion for advanced prostate cancer.

5.1 Introduction

The goal of the conventional clustering algorithms is to classify objects based on a measurement of similarity. Several measurements of the similarity were proposed in the literature to enhance the clustering quality. The objects in each cluster must share specific and unique characteristics which isolate them from objects in other clusters, intruders, and noise.

The measurement of similarity tends to be uniform as the space dimensionality increases. In this case, all objects become equally-spaced which makes each object similar to all other objects, and thus, no proper clustering can be done. This behavior is known as “curse of dimensionality” and unfortunately, it is the case in most real-world datasets. However, it is possible that a group of objects have more similarity in one dimension or in a smaller subspace than their similarity in the whole space. This leads to the subspace clustering methods which search for subspaces where the objects are similar and distinguishable in some sense.

Many methods were proposed during the last decade. These methods consist mainly of two steps: first, they test all one dimensional features to determine their ability to cluster the objects in distinct groups. The capable features are then selected to form the set of “candidate features”. Second, all different combinations of the candidate features are tested to build up the sought-after subspace sequentially.

The complexity of the second step, in most methods, depends exponentially on the number of selected features. This number grows linearly with dimensionality of the space and that makes the problem very ill-conditioned for most real-world datasets. The number of the selected features can be controlled by the level of the similarity measurements that is used in the first step. Requiring a rigid similarity measurement yields smaller sets of candidate feature but also, it discards some significant features. On the other hand, allowing lenient measurement of similarity selects all or most of the significant features at the expense of the set’s size. The trade-off between less

complexity and higher accuracy favors the reduction of the complexity since the process becomes infeasible if the size of the candidate features' set exceeds a certain limit. The size of the candidate features' set must be reduced, and in the same time all significant features must be maintained in order to have an effective clustering method.

5.2 Related Work

Six methods are extensively used in the literature [82-87]. The first published method in subspace clustering criteria is CLIQUE [82]. Each dimension is portioned into equally spaced non-overlapping intervals. The measurement of density is used to evaluate the features. The number of objects that fall in each interval determines the density. An interval is declared dense if its density exceeds a minimum threshold, and a feature is considered as a candidate if it includes at least one dense interval. ENCLUS [83] selects the candidate features based on their entropy. It measures the entropy of each feature and adds it to the candidate set if its entropy is less than a threshold. The feature whose entries carry the same value must be excluded since the entropy is equal to zero whereas its result is meaningless. MAFIA [84] represents a more significant improvement to CLIQUE by suggesting an unequally-spaced grid for each feature. It starts by checking the histogram of the entries of each feature in a very fine grid. Then it combines successive intervals if the difference between the values of their populations is less than a threshold.

Other methods extend the density concept by introducing the density-connected approaches like SUBCLU [85]. It requires a minimum threshold of objects to form a dense unit and a maximum distance threshold to guarantee the connectedness. If an entry point has enough points in its proximity, it will be declared a core point, and the feature will be selected. Any feature that does not include at least on core point will be excluded. PreDeCon [86] is a modified version of SUBCLU which requires the variance of the entries that fall in the neighborhood of an entry to be less than a

maximum threshold. K-mean is suggested by FIRES [87] as well as any other practical measurement of similarity to be used. After clustering the entries of all features, the average number of entries in each cluster is calculated. Any cluster that includes entries less than 25% of their mean is excluded. Finally the T-test was used in [88] to cluster the same data that we will be testing at this work.

5.3 Segregated-Based Clustering Algorithm SBC

We got the data set ζ from [87]. ζ is a set a arrays that are processed by MIS as described in chapter 2. It includes samples from 142 individuals where 71 of them are autistic (AU) and 71 are typically developing (TD). The data covers 384,432 positions of the genome, ($\zeta \subseteq \Re 384{,}432$). The set of all features is $F = \{F_1,\ldots,F_{384,432}\}$. Presumably, there is at least one hidden subspaces comprising k features ($k << 384{,}432$) where the objects cluster in it in a meaningful way. We are seeking the subspace that clusters the objects into the purest possible clusters of AU or TD labels.

We present SBC, which follows the customary bottom-up approach to build up the hidden discriminative subspace. It allows using any of the similarity measurements presented in the previous section. It employs a quality factor to remove insignificant features from the candidate features set. The quality factor is a percentage value representing the purity of labels in each cluster. The conventional clustering analyses are driven by the similarity measurement whereas SBC is driven by the similarity measurement that provides semi-pure clusters. The reduction of the complexity is considerable while the accuracy is maintained.

After acquiring the candidate features set, which should be the smallest efficient possible set, a conventional forward-feature-selection approach is applied to build up the required subspace. It tests all possible cubic subspaces to choose one feature to be added to the required subspace. The same quality factor of segregating purity is applied at each step.

We applied the methods in [82-87] on our data to create the candidate features set. We then added the segregation quality factor to show the large reduction of the size of the candidate features' sets. Then, we recursively built up a 13-D subspace which provided the purest possible segregation we achieved. Finally, we generated random data and mapped it to the formed subspace to assure the result's validity, and we also applied the "leave-one-out" approach to test its stability.

SBC assumes that the number of clusters, η, in the hidden subspace is known a priori. This number should be large enough to cover the heterogeneity in the object's groups and the centroids are allowed to overlap at some features. SBC also assumes that all clusters have the same dimensionality, k, which is unknown at the beginning. The customary two-step approach is represented in the following sections.

5.3.1 Collecting the set of candidate features:

SBC allows using any of the techniques mentioned in section 5.2 to assemble the candidate features set. Then, it tests every selected feature to decide whether to save it or to ignore it. This test measures the purity of the labels AU and TD in each dense interval, unit, or cluster found in any 1-D feature using any of the previously mentioned methods. The purity is measured as:

$$purity = \frac{1}{142}\sum_{i=1}^{\eta} \max(TD_i, AU_i) \qquad (5.1)$$

TDi: the number of TD labels in cluster i

AUi: the number of AU labels in cluster i

Any feature that does not satisfy the purity requirement is excluded regardless of how dense it is. It is of note that the maximum value of purity is 1 and the minimum value is 0.5. The number of clusters was varied from 3 to 12 and all mentioned methods were tested. We chose a supervised k-mean criterion in our test.

5.3.2 Bottom-up approach:

After selecting the candidate features, SBC works as the following: It starts with Purity = 0.5 and with an empty subspace. Then, it tests all possible 3-D subspaces to create a *promising* features' set of size 3. For each 3-D subspace, SBC projects the data into k clusters and computes the segregation purity as explained in Eq. 5.1. The 3 features that exhibit the highest purity are named F_1, F_2, and F_3.

To add one more dimension to the synthesized subspace, SBC tests all 4-D subspaces that consist of the union of F_1, F_2, or F_3 with any additional 3-D subspace from the set of the left candidates. The segregation purity is measured at each permutation. When the 4-D subspace that exhibits the highest purity is found, and if the purity level of this 4-D subspace is higher than the purity level provided by the previous 3-D subspace, then the corresponding feature, F_i (i = 1,2, or 3), is added to the synthesized subspace. The other three features that form this 4-D subspace with F_i will be named F_1, F_2, and F_3, and the same step will be repeated again. The technique continues repeating this step until the purity saturates or starts to decrease, then it adds the set of $\{F_1, F_2, F_3\}/\{F_i\}$ to the synthesized subspace and stops. Figure 5.1 illustrates the flowchart SBC.

The computational cost of finding the candidate features' set is always linearly proportionate to the whole space dimensionality and to the number of objects. But the complexity of building the required subspace is proportionate to the size of the candidate features' set exponentially or cubically as shown in table 5.1.

5.4 Experimental Results of Autism

The 142 objects are located in a 384,432-dimension space. We applied the methods [82-87] to generate different sets of candidate features. The size of all candidate features' sets for all methods was reduced to ~25% after considering the purity restriction. Any feature whose purity is less than 0.56 was excluded.

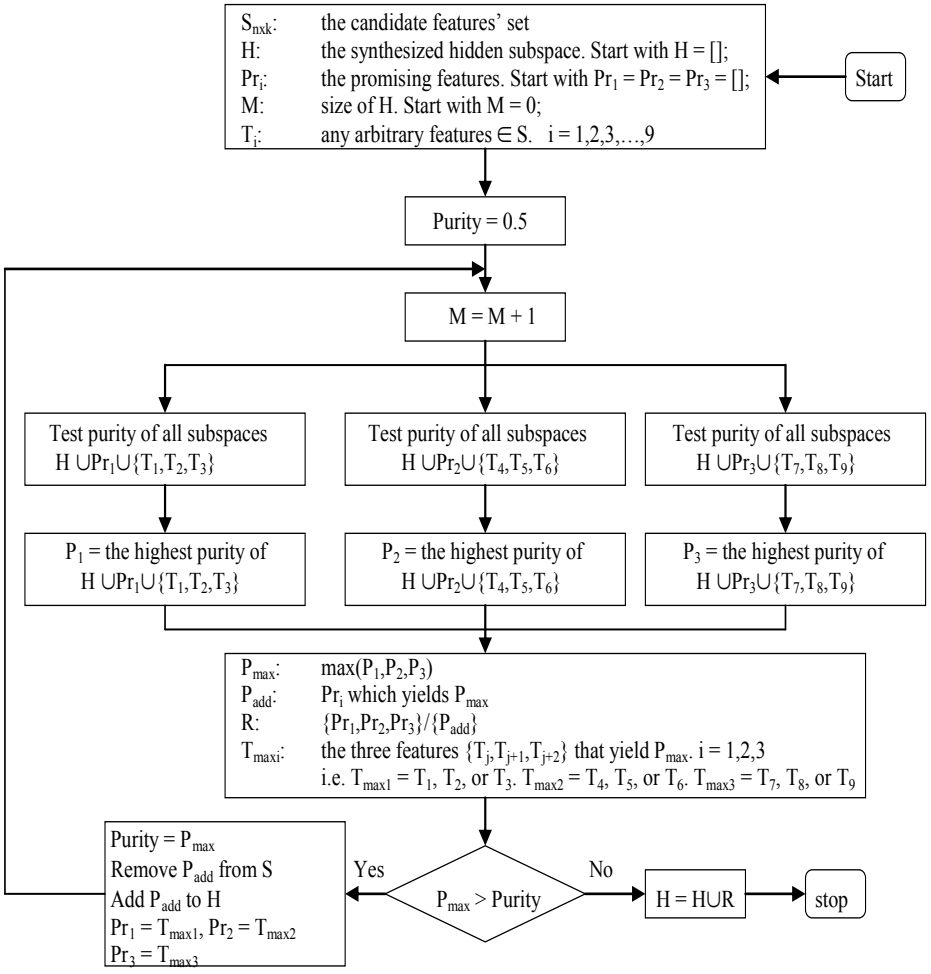


Figure 5.1: Flowchart of SBC

Technique	Complexity	Remarks
CLIQUE	$O(mk + C^k)$	m: number of object k: hidden subspace size C: is a constant > 1 d: the whole space dimension $d >>> k$
ENCLUS	$O(mk + C^k)$	
MAFIA	$O(mk + C^k)$	
SUBCLU	$O(mk + C^k)$	
FIRES	$O(mk + C^k)$	
PreDeCon	$O(d.k^2)$	
SBC	$O(k^3)$	

Table 5.1: Complexity of different techniques

For SBC, an unsupervised k-mean clustering algorithm was implemented for each feature. The number of clusters, k, was ranged from 3 to 12 with a fixed value of k for all features in every implementation. After finding the centroids of k-clusters, all 142 data points are mapped into the nearest clusters using the Euclidian distance. Cluster's purity was measured for each feature and the same restriction was applied. The results of all methods before and after considering the purity restriction are shown in table 5.2.

Technique	Size before purity restriction	Size after purity restriction	Complexity order
CLIQUE	1202	339	$O(10^{100})$
ENCLUS	2241	-	$\rightarrow \infty$
MAFIA	2332	577	$O(10^{167})$
SUBCLU	1082	312	$O(10^{93})$
FIRES	713	241	$O(10^{72})$
PreDeCon	5478	1432	$O(10^{10})$
SBC	-	41	$O(10^{5})$

Table 5.2: The size of the candidate features' set of different tecniques

Cluster	**T-test Technique**		**SBC**	
	TD/AU	status	TD/AU	status
1	35/6	85%TD	66/16	80.5% TD
2	9 /38	80% AU	3/5	62.5% AU
3	5/16	76% AU	0/24	100% AU
4	21/12	64% TD	0/11	100% AU
5	-	-	2/15	88.2% AU
Purity	77%		85%	

Table 5.3: Comparative clustering results

It is clear from table 5.2 that the second step of the bottom-up approach of all methods, except SBC, can not be carried out at the given dataset due to the huge dimensionality. The exponential complexity for selecting 41 candidate features blows up to infinity. It is also noticeable that k-mean provides the smallest set. Under cubical growth of complexity, the computational cost increases 200-fold when the candidate features' set's size increases from 41 to 241 which is the minimum size any other method can provide. The same data were tested in [88] and a comparison with their results is presented in table 5.3.

5.4.1 Results and Conclusion

The choice of using 5-clusters was found to be the best. The synthesized subspace consists of 13-dimensions and it is found to be the most discriminative subspace with 4 AU clusters and one TD cluster. The purity is 85.2% (121 individuals are classified correctly) with two 100% pure clusters containing only autistic individuals. See table 5.3. It is of note that the TD cluster has the least norm which corresponds to the least variation of DNA copy number as shown in figure 5.2.

More interesting, it is found that the objects were clustered based on the source of their DNA samples! All DNA samples in the fifth cluster were obtained from blood cells except one individual whose sample was obtained from transformed cells. All AU individuals in cluster # 2 and 3 were obtained from transformed cells except 2 individuals in cluster # 3 were obtained from the whole blood cells. Cluster # 4 is a mix with 7 AU individuals from the whole blood cells and 4 from transformed cells. The information about the source of the DNA was unknown during the process and it proves the robustness of our method as it can capture the slight differences of the readings of the same population to divide them into two groups without misidentifying them.

80% of the individuals that are classified as TD are typically developing in real. 92% of individuals that are classified as autistic are autistic in real. 77.5% of the AU individuals were classified AU, and 93% of TD individuals were classified TD.

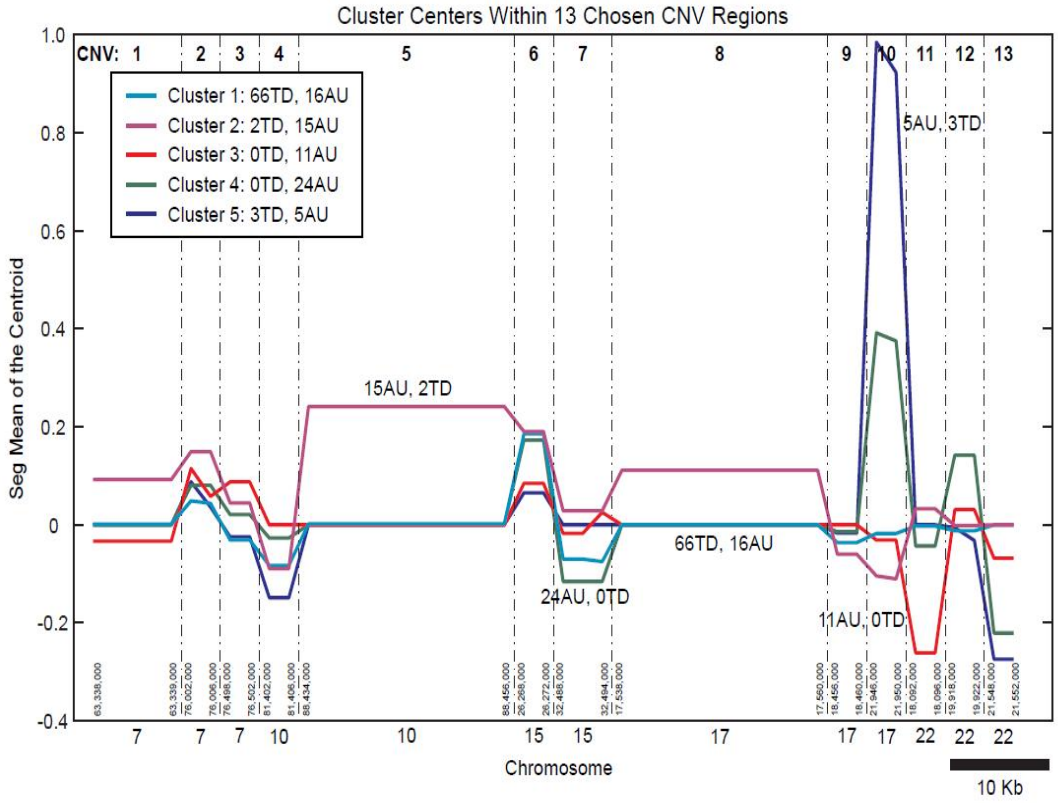


Figure 5.2: Centroids of 5 Clusters in 13-D sup-space

It is important to emphasize that the result of SBC is not unique. It provides multiple synthesized subspaces and they share the same purity and the majority of their features but they are different in their stability and validity. The "leave-1-out" test was applied to measure the controids' stability. The subspace whose centroids have the least norm of variations was chosen. We also created a set of 1000 13-D arrays with completely random values to test the validity of our subspace. The arrays' random values are independent and they vary uniformly from the minimum to the maximum values of the range at each dimension. A Euclidian distance = 1 is used as a maximum threshold to permit a point to belong to a cluster. All true labels fell into their clusters with the required distance whereas no random data points were close enough to be accepted by any cluster.

5.5 The Association of DNA Copy Number Variations with Prostate Cancer Therapy

Healthy prostate tissue is stabilized by specific required androgens that bind to specific receptors known as androgen receptors AR [89]. The locus of AR covers about 110kkDa of the male X-chromosome. The growth of prostate cancer is contingent to the normal activity of the AR locus, and therefor, it can be controlled by using androgen depletion [90]. Androgen depletion restrains the normal activities of the AR which is needed for the growth of the disease.

A new phenomenon has been noticed recently where the disease becomes out of control even with the continuation of using the same androgen depletion. We investigated the copy number variation on the AR locus in two groups of male individuals. The datasets are publicly available at http://www.ncbi.nlm.nih.gov/geo/ with GEO accession numbers GSE18333 and GSE14996. The first group responds to the therapy (CWR22Pc cells) while the second group does not (CRPCa cells). Large copy number variations are detection solely in the AR locus of the individuals of the second group. We concluded in [92] that the excessive duplication (>50 folds) in the CRPCa cells altered the sequence of the AR locus and created AR isoforms which do not respond to the therapy and at the same time enhances the growth of the disease.

5.6 Conclusion

We presented a novel algorithm to data-mine in huge dimensional datasets to discover patterns of similarities among different objects in an efficient time. We applied the algorithm at autistic samples and advanced prostate cancer patients. We discovered patterns of copy number variation that can indicate or predict the existence of the phenotype in any other sample.

Chapter 6

Conclusion and Future Work

The field of analyzing the human genome and detecting the abnormalities in its structure remains one of the most promising areas in the way to a full understanding of the human genome and its functions. Several diseases and phenotypes have been proven to be directly effected by certain types of variation of the genome whether in its structure as copy number or in its production as genes and protein. The field is certainly very promising and it has a great deal of potential in the treatment and diagnosis of several diseases. We hope that this work will provide a useful contribution to the field and we hope that it will inspire researchers to accumulate the knowledge necessary to complete the human genetic map.

We discussed three main areas in this work and we provide here some potential work that may improve them.

6.1 Two Channel Approaches

We presented a comprehensive experiment of the two-channel approaches. We presented four models based on the Band-Pass Wavelet Transform, Uncovered Markov Model, Truncated Likelihood Ratio Test, and Minimum Interval Score. We also discussed the reproducibility of the receiver operating characteristic curves (ROC). We ran an experiment to test the stability of the ROC curves and we proved that they are stable at a relatively narrow band where the sensitivity is high and the false alarm rate is low. This band is the preferred band for any experiment and the only band that provides acceptable results.

Several parts of the two-channel approaches need to be addressed:

1. The distribution of the observations is still not exactly known. Almost all the existing algorithms, as well as our models, assume the Gaussian process. However, the exact distribution of the microarrays observations carries a slightly heavier tails than the Gaussian distribution and that generates an amount of false rate much higher than what is expected based on the theory. We specifically emphasized in the sensor network's results that the last 15% of each side of the distribution do not fit with the Gaussian model. The closest model to the distribution, other than the Gaussian, is the Logistic distribution. However, it also does not fit the observations perfectly. A new model for the distribution may provide a significant improvement to the field to reduce the generated error.
2. The reproducibility of the ROC curves using different data is an intriguing subject that needs to be addressed fully and thoroughly. We ran an experiment of a cross-validation approach to prove it, but we believe that the topic can be addressed better if more datasets from different labs and platforms are used.

6.2 Single-Channel Approach

We presented the first single-channel approach in the field. It analyzes the array entirely within itself without using a reference to make a comparison. Its results are quantified in numbers of DNA copies as opposed to the two-channel approaches where the results are comparative which only indicates if the test has more or less copies than the reference. We believe some parts of the model can be improved:

3. The Universal Threshold Adjustment to remove the bias of the imperfect scanner still requires some improvement. This bias is significant and persistent in all microarrays and it is shocking that a very little effort has been dedicated to address this matter in the literature. We showed that Local Adjustment is better in removing the bias but it alters the real variation of the DNA copy number. The UTA removes approximately 80% of the bias while preserving the real variation without alteration. An improvement to this performance is definitely desired.

4. We investigated the saddle-point approximation under the Bayesian criteria and we proved its accuracy and efficiency in analyzing the non-homogenous environment of the microarrays. The approximation is also applicable under the Neyman-Pearson theory but we have not explored this part yet. Neyman-Pearson theory provides the Uniformly Most Powerful Test to analyze the microarrays. Employing the saddle-point approximation in the analysis will provide a strong tool that is accurate, reliable, and time efficient.
5. We explored the variability and the stability of the human genome. We analyzed 1258 samples of 19 different populations from several parts of earth. However, this number is infinitesimal to generalize conclusions about the 6.97 billion humans alive now. The margin of error is considerably higher than reaching reliable conclusions that can fit all human beings. More samples are needed to get results with a narrow margin of error with a higher level of confidence about the validity of the conclusion.

6.3 Subspace-Clustering Algorithms

We presented a new model for data-mining the human genome. The signals of the human genome have an extremely high dimensionality which require a significant reduction of the computational load to be analyzed. The goal of this model is to reveal certain patterns of copy number variation in special groups. We detected highly correlated variant regions in autistic and prostate cancer samples. The applications of this model are countless especially with the availability of tens of thousands of DNA samples on the National Center for Biotechnology Information's website. Applying the model to a special group whose individuals share a certain phenotype might provide a biological signature of that studies phenotype. The biological signature can be employed to detect the existence of a disease or the susceptibility of endorsing it in the future. It also will be very crucial in preventive healthcare and early treatments.

Reference

[1] Redon, R. et al. Global variation in copy number in the human genome. Nature 444, 444–454, 2006.

[2] Freeman JL, et al. Copy Number Variation: New Insights in Genome Diversity. Genome Research; 16:949–961, 2006.

[3] Balciuniene J, et al. Recurrent 10q22–q23 deletions: a genomic disorder on 10q associated with cognitive and behavioral abnormalities. Am J Hum Genet; 80:938–47, 2007.

[4] Hert D.G., Fredlake C.P., Barron A.E. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis; 29:4618-4626, 2008.

[5] A. K. Alqallaf and A. H. Tewfik. DNA copy number detection and Sigma filter. GENSIPS, pp. 1–4, June 2007.

[6] Hu J, Gao J-B, Cao Y, Bottinger E, Zhang W. Exploiting noise in array cgh data to improve detection of dna copy number change. Nucleic Acids Res.; 35:e35, 2007.

[7] Fitzgerald TW, Larcombe LD, Le Scouarnec S, Clayton S, Rajan D, Carter NP, et al. aCGH-Spline: An R package for aCGH dye bias normalisation. Bioinformatics; 27:1195–1200, 2011.

[8] Hsu, L., et al. Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6211–226, 2005.

[9] Khojasteh M, Coe BP, Shah S, et al. A Novel Algorithm for the Analysis of Array CGH Data. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006.

[10] Huang H, et al. Array CGH data modeling and smoothing in stationary wavelet packet transform domain. BMC Genomics; 9:S2-S17, 2008.

[11] M.S. Islam. Periodicity, Change Detection and Prediction in Microarrays. Ph.D. Thesis, University of Western Ontario, 2008.

[12] NEXUS COPY NUMBER Software. Version 4.1. 2009.

[13] Andersson R, Bruder CEG, Piotrowski A, Menzel U, Nord H, Sandgren J, Hvidsten TR, de Sthl TD, Dumanski JP, Komorowski J. A segmental maximum a posteriori approach to genome-wide copy number profiling. Bioinformatics; 24:751–758, 2008.

[14] Rueda OM, Diaz-Uriarte R. A flexible statistical method for detecting genomic copy-number changes using Hidden Markov Models with reversible jump MCMC. COBRA preprint series, 2006.

[15] NimbleScan Sofrware User's Guide. Version 2.6, 2010.

[16] Akaike H. A new look at the statistical model identification. IEEE Trans. Auto. Control 19: 716–723, 1974.

[17] Schwarz, G. Estimating the dimension of a model. Ann. Statist. 6461–464, 1978.

[18] Lebarbier E. Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection. Signal Processing 2005.

[19] Lavielle M. Using penalized contrasts for the change-point problem. Signal Processing 85.8, 2005.

[20] Tibshirani R, Wang P. Spatial smoothing and hotspot detection for CGH data using the fused lasso. Biostatistics; 9:18-29, 2008.

[21] Charalampos E. Tsourakakis1 , David Tolliver1 , Maria A. Tsiarli2 , Stanley Shackney3 , Russell Schwartz. CGHTRIMMER: Discretizing noisy Array CGH Data. Cornell University Library arXiv:1002.4438v1, 2010.

[22] Erdman C, Emerson JW. A fast Bayesian change point analysis for the segmentation of microarray data. Bioinformatics, 24:2143-2148, 2008.

[23] Wang, P., et al. A method for calling gains and losses in array CGH data. Biostatistics 645-58, 2005.

[24] Morganella S, Cerulo L, Viglietto G, Ceccarelli M. VEGA: variational segmentation for copy number detection. Bioinformatics, 26(24):3020-3027, 2010.

[25] Price TS, et al. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. Nucleic Acids Res.; 33:3455–3464, 2005.

[26] Pollack J, Perou C, Alizadeh A, Eisen M, Pergamenschikov A, Williams C, Jeffrey S, Botstein D, and Brown P. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genet., 23, 41–46, 1999.

[27] Rabiner L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2): 257–286, 1989.

[28] Fridlyand, J., et al. Hidden Markov models approach to the analysis of array CGH data. J. Multivariate Analysis, 90132–153, 2004.

[29] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics; 5:557–572, 2004.

[30] Picard, F., et al. A statistical approach for array CGH data analysis. BMC Bioinformatics 627, 2005.

[31] GILL, P., MURRAY, W. AND SAUNDERS, M. Users guide for SQOPT 5.3: a Fortran package for large-scale linear and quadratic programming. Technical Report. Stanford University, 1999.

[32] Barry,D. and Hartigan,J.A. A Bayesian analysis for change point problems. J. Am. Stat. Assoc., 88, 309–319, 1993.

[33] Lipson D, Aumann Y, Ben-Dor A, Linial N, and Yakhini Z. Efficient calculation of interval scores for DNA copy number data analysis. J. Comput. Biol.; 13:215–228, 2006.

[34] Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 21: 4084–4091, 2005.

[35] Lai WR, et al. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics; 21(19):3763–70, 2005.

[36] Hehir-Kwa J, Egmont-Petersen M, Janssen I, Smeets D, Geurts van Kessel A, Veltman J. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res.; 14:1–11, 2007.

[37] Dellinger AE, et al. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res.; 38:e105, 2010.

[38] Tsuang D.W, Millard S.P, Ely B, Chi P, Wang K, Raskind W.H., Kim S, Brkanac Z. and Yu C.E. The effect of algorithms on copy number variant detection. PLoS ONE; 5:e14456, 2010.

[39] Grayson BL, Aune TM. A comparison of genomic copy number calls by Partek Genomics Suite, Genotyping Console and Birdsuite algorithms to quantitative PCR. BioData mining 4: 8, 2011.

[40] Koike, A, Nishida D. Yamashita, and K. Tokunaga. Comparative analysis of copy number variation detection methods and database construction. BMC Genetics. 12:29, 2011.

[41] Lin, P., Hartz, S. M., Wang, J. C., Krueger, R. F., Foroud, T. M., Edenberg, H. J., et al. Copy number variation accuracy in genome□wide association studies. Hum Hered, 71(3), 141□147, 2011.

[42] ECKEL-PASSOW, J. E., ATKINSON, E. J., MAHARJAN, S., KARDIA, S. L. R., AND DE ANDRADE, M. Software comparison for evaluating genomic copy number variation for affymetrix 6.0 snp array platform. BMC Bioinformatics 12, 220, 2011.

[43] Alqallaf A, Tewfik A, and Selleck S. Genetic variation detection using maximum likelihood estimator. IEEE International Workshop on Genomic Signal Processing and Statistics, 2009.

[44] Shen F, Huang J, Fitch KR, Truong VB, Kirby A, Chen W, Zhang J, Liu G, McCarroll SA, Jones KW, et al. Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. BMC Genet.; 9:27, 2008.

[45] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F, A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc 99: 909–917, 2004.

[46] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–211, 1998.

[47] Lucito, R., West, J., Reiner, A., Alexander, J., Esposito, D., Mishra, B., Powers, S., Norton, L. and Wigler, M. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res., 10, 1726–1736, 2000.

[48] Bignell GR, Huang J, Greshock J, et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res; 14: 287–95, 2004.

[49] H. Vikalo, B. Hassibi, and A. Hassibi, A statistical model for microarrays, optimal estimation algorithms, and limits of performance, IEEE Transactions on Signal Processing, Special Issue on Genomics Signal Processing, vol. 54, No. 6, Jun. 2006, pp. 2444-2455.

[50] Wu, L., D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje, and J. Zhou. Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl. Environ. Microbiol. 67:5780-5790, 2001.

[51] Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. Proceedings of the National Academy of Sciences of the United States of America; 103:12457–12462, 2006.

[52] L. Wan, W.J. Fu, M. Deng, and M. Qian. A method to correct systematic bias in Affymetrix snp arrays. Proceeding of The International Conference on BioMedical Engineering and Informatics, pp. 442-6, 2008.

[53] D. Seale and S. W. Davies, Stochastic model of DNA microarray, in Proceeding of IEEE Special Top. Conf. Mol., Cell., Tissue Eng., pp. 113–114, 2002.

[54] Rigaill G, Hupe P, Almeida A, La Rosa P, Meyniel J, Decraene C, Barillot E. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics; 24:768–774, 2008.

[55] Diskin S, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res.; 36:e126, 2008.

[56] Wan, L, Xiao Y, Chen Q, Deng M, and Qian M. The Biases of Copy Numbers from Affymetrix SNP Arrays and Their Corrections. Communications in information and systems 10.2 :83, 2010.

[57] Vikalo H, Hassibi B, Hassibi A. ML Estimation of DNA Initial Copy Number in Polymerase Chain Reaction (PCR) Processes. Acoustics, Speech and Signal Processing. Vol 1, April 2007.

[58] H. Vikalo, B. Hassibi, M. Stojnic, and A. Hassibi. Modeling the kinetics of hybridization in microarrays, IEEE Intern. Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland, 2007.

[59] Hooyberghs J, Van Hummelen P, Carlon E. The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. Nucleic Acids Research 37: e53, 2009.

[60] Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res.; 13:2291–2305, 2003.

[61] Chuang, H.Y., Tsai, H.K., Kao, C.Y. Optimal designs for microarray experiments. 7th International Symposium on Parallel Architectures, Algorithms, and Networks, Hong Kong, China, pp. 619–624. IEEE Computer Society, Los Alamitos, 2004

[62] Palmer J. The Analysis of Oligonucleotide Microarray Data at the Raw Image Level and the Probe Level. Ph.D. Thesis, Carnegie Mellon University, 2005.

[63] Hu P, et al. Integrative analysis of gene expression data including an assessment of pathway enrichment for predicting prostate cancer. Cancer Inform.; 2:289-300, 2006.

[64] Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, et al. A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics 12: 33–50, 2011.

[65] Zhang L, et al. A model of molecular interactions on short oligonucleotide microarrays. Nat. Biotechnol.; 21:818–821, 2003.

[66] Wei Wei, Lin Wan, Minping Qian, and Minghua Deng. The Implementation and Application of the Microarray Preprocessing Generalized PDNN Model. 2nd International Conference on Biomedical Engineering and Informatics, 2009.

[67] Rocke, D., and Durbin, B. A model for measurement error for gene expression arrays. J. Comput. Biol. 8, 557–569, 2001.

[68] Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65: 6071–6079, 2005.

[69] Banneheka, B.M.S.G., Ekanayake, G.E.M.U.P.D. A New Point Estimation for the Median of Gamma Distribution. Viyodaya Journal of Science, Vol 14. pp. 95-103, 2009.

[70] Khojasteh M, Lam WL, Ward RK, MacAulay C. A stepwise framework for the normalization of array CGH data. BMC Bioinformatics 6:274, 2005.

[71] Wilson, D.L., Buckley, M.J., Helliwell, C.A. and Wilson I.W. New normalization methods for cDNA microarray data. Bioinformatics 19: 1325–1332, 2003.

[72] Song JS, et al. Model-based analysis of two-color arrays (MA2C) Genome biology; 8:R178, 2007.

[73] Lisovich, A., Chandran, U. R., Lyons-Weiler, M. A., LaFramboise, W. A., Brown, A. R., Jakacki, R. I., Pollack, I. F., & Sobol, R. W. A novel SNP analysis method to detect copy number alterations with an unbiased reference signal directly from tumor samples. BMC Med Genomics, 4:14, 2011.

[74] Halldorsson B, Gudbjartsson D. An algorithm for detecting high frequency copy number polymorphisms using SNP arrays. Journal of Computational Biology, 18:955-66, 2011.

[75] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics; 12:480, 2011.

[76] Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, et al. CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics; 7:83, 2006.

[77] S.A. Aldosari and J.M.F. Moura, Saddlepoint approximation for sensor network optimization, ICASSP, vol. 4., pp. 741–744, Mar. 2005.

[78] Bengtsson H, Wirapati P, Speed TP. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWide SNP 5 &6. Bioinformatics 25: 2149–2156, 2009.

[79] Clevert D-A, Mitterecker A, Mayr A, Klambauer G, Tuefferd M, Bondt AD, Talloen W, Göhlmann H, Hochreiter S. cn.FARMS: A Latent Variable Model to Detect Copy Number Variations in Microarray Data with a Low False Discovery Rate. Nucleic Acids Res.; 39:e79, 2011

[80] R. Lugannani and S. Rice. Saddle-point approximation for the distribution of the sum of independent random variables. Advanceds in Applied Probabilit, 12:475–490, 1980.

[81] R. Agrawal, J. Gehrke, D. Gunopulos, and. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the SIGMOD Conference, Seattle, WA, 1998.

[82] C. H. Cheng, A. W.-C. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, pages 84-93, 1999.

[83] S. Goil, H. Nagesh, and A. Choudhary. MAFIA: Efficient and scalable subspace clustering for very large data sets. Technical Report CPDCTR-9906-010, Northwestern University, 1999.

[84] K. Kailing, H.P. Kriegel, and P. Kroger. Density-connected subspace clustering for high-dimensional data. In Proceedings of the 4th SIAM International Conference on Data Mining (SDM), Orlando, FL, 2004.

[85] H.P. Kriegel, P. Kroger, M. Renz, and S. Wurst. A generic framework for efficient subspace clustering of high-dimensional data. In Proceedings of the 5th International Conference on Data Mining (ICDM), Houston, TX, 2005.

[86] C. Bohm, K. Kailing, H.P. Kriegel, and P. Kroger. Density connected clustering with local subspace preferences. In Proceedings of the 4th International Conference on Data Mining (ICDM), Brighton, U.K., 2004.

[87] A. Alqallaf, A. Tewfik, P. Krakowiak, F. Tassone, R. Davis, R. Hansen, I. Hertz-Picciotto, I. Pessah, J. Gregg and S. Selleck. Identifying Patterns of Copy Number Variatnts in Case-control Studies of Human Genome Disorders. In the proceedings of Genomic Signal Processing and Statistics conference, Gensip, Minneapolis, MN, 2009.

[88] Abdullah K. Alqallaf. Signal Processing Techniques and Statistics for the Analysis of the Human Genome. Ph.D. Thesis, University of Minnesota, 2009.

[89] Heemers HV, Tindall DJ. Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex. Endocr Rev; 28:778-808, 2007.

[90] Taplin ME. Drug insight: role of the androgen receptor in the development and progression of prostate cancer. Nat Clin Pract Oncol ;4:236-44, 2007.

[91] Chen Y, Clegg NJ, Scher HI. Anti-androgens and androgen-depleting therapies in prostate cancer: new agents for an established target. Lancet Oncol; 10:981-91, 2009.

[92] Li Y, Alsagabi M, Fan D, Bova GS, Tewfik AH, Dehm SM. Intragenic rearrangement and altered RNA splicing of the androgen receptor in a cell-based model of prostate cancer progression. Cancer Res.; 71:2108–17, 2011.

[93] Pinto D, Darvishi K, Shi X, Rajan D et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 8;29(6):512-20. PMID, 2011.